Bias and equivalence of psychological measures in South Africa

Meiring, D.

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Deon Meiring

Bias and Equivalence of Psychological Measures in South Africa

BIAS AND EQUIVALENCE OF PSYCHOLOGICAL MEASURES IN SOUTH AFRICA

BIAS AND EQUIVALENCE OF PSYCHOLOGICAL MEASURES IN SOUTH AFRICA

PROEFSCHRIFT

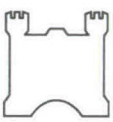
ter verkrijging van de graad van doctor aan de Universtiteit van Tilburg, op gezag van de rector magnificus, prof.dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Ruth First zaal van de Universiteit op dinsdag 20 februari 2007 om 14.15 uur

door

Deon Meiring
geboren op 21 juni 1967 te Carletonville, Zuid-Afrika

Promotors: Prof.dr. A.J.R. van de Vijver
Prof.dr. S. Rothmann

Publisher: Labyrint Publication
Postbus 334
2984 Ax Ridderkerk
The Netherlands
Tel: 0180 – 463962

Printed by:  Offsetdrukkerij, Ridderprint B.V., Ridderkerk

ISBN-10: 90-5335-111-6
ISBN-13: 978-90-5335-1116

Acknowledgements

This PhD project was a steep climb for me and it felt like climbing Mount Everest not knowing if I would ever reach the summit. It started five years ago when I was faced with the situation that psychological tests were banned in South Africa. The Employment Equity Act No. 55 of 1998 (section 8) prohibited the use of psychological tests in South Africa, unless it complied to stringent criteria. My need to overcome this situation began with the scouting of a champion leader, a person who was at the top of his game in this field of psychometrics and cross-cultural psychology and who could lead an expedition. I tracked Prof Fons van de Vijver down in the late autumn of 2001 in Tilburg with a fellow climber Prof Ian Rothmann who joined my expedition to explore this challenge. Our first encounter with Tilburg and Fons was a much anticipated one. Fons had booked us in at the Grand Central Hotel in Tilburg. The hotel was central but there was nothing grand about it. Later we were upgraded to the Auberge Hotel. Fons started out with his very famous explanations of how bias and equivalence statistics worked. Ian and I did not understand much about this and so a great journey started.

Over a period of five years the three of us and fellow collaborators, Murray Barrick, Paul Sackett and Deon de Bruin started with the planning for the climbing of this mountain. It started out with setting up base camp and over the next years various expeditions were carried out to camp 1 (article 1- twelve months of work), camp 2 (article 2 - eleven months of work), camp 3 (article 3 - eighteen months of work), camp 4 (article 4 - seven months of work) with the final ascend in 2006 to the summit. Ian and I visited Tilburg eight times to work on the project and Fons, who also became a special professor at the University of North West (Potchefstoom Campus) in 2004, visited South Africa seven times. Who will ever forget the encounter that Fons, Lona and I had with an elephant bull in the Kruger National Park or Fons and Ian's encounter with a python snake on their way to the Mafikeng Campus.

Communication during this climb was important and over the years up to 8 000 emails were sent between the expedition team

members, the bulk of this being to and from Fons, who had this amazing ability to reply within the same day or even the same hour!. Later on in the project we switched over to Skype which enabled us to be in more direct contact.

Fons, I want to thank you for you guidance and mentorship during this journey. You always knew what the final destination would be. We shared great times and moments together, who will forget our trip to China in 2004 where we presented our first set of results at the International Association for Cross-Cultural Psychology (IACCP) conference. What I will always admire about you is your strong leadership, your vision, your perseverance and tenacity to never give up when faced with a problem. Your helpfulness, guidance, kindness, gentleness, sensitivity and your wisdom will always stay with me. As I continue with my career, I pray that I will be the same role model to my students as you were to me, after all I have learned from the best in the world.

Then, to Ian Rothmann who I knew from my Bachelors days (1986) at the North-West University (Potchefstroom Campus). You where there when we started this climb and you were there when we summited. Someone who is a true friend in good or bad times, someone that you can always rely on, someone who always helped, despite his own workload, someone who will wake up at one o'clock in the mornings many times to have his input ready for the next morning at seven o'clock. Yes, Ian you are a true friend, I will always cherish our special relationship. Some of our greatest projects and schemes were conceptualized in Tilburg and who will ever forget our wonderful times we had at "Het Dorstige Hert" on the tune of "Suiker Bossie" by Stef Bos.

Then there were many other people at Tilburg University who were good to me, to my fellow PhD students, Otmane, Dianne, Judit, Seger, Symen, Maike and Irina thanks for your support and encouragement. A special thanks to Saskia who assisted me at the very end with the formatting of my document. Rinus, thank you for the editing of my document, you did a great job. To Robbie who always assisted me with my accommodation arrangements when I traveled to Tilburg, thank you. To Suzette Kielblock and her children, thanks for the lovely times we spent together in Utrecht.

Also to the Babylon Center for its financial support during my visits. And then last but not the least to Ype Poortinga who showed a keen interest in my work and South Africa, thanks for your encouragement and kind words when I was really low and exhausted with the climb.

The South African Police Services, Psychological Services who gave me the opportunity to conduct this research. Thanks to my colleagues at work and especially to Director Anton Grobler who always supported me, without your encouragement I would have not reached the summit.

Then to Steve who died tragically in 2002, you always believed in me from the start to reach the top. You bought me my first equipment (laptop) to climb the mountain and you also sponsored my first trip to Tilburg with your voyager miles. I will also tribute this achievement to you. Also to Amelia who showed a keen interest into my progress, thank you.

Then to the most important people close to me, my dad who passed away in 2004 during my journey. It was always your vision to see me achieve and that I did indeed. Thank you for your part that you played in my live. My mother who were always proud of my achievements, thank you for believing in me.

Then to my life companion Ilse and our two beautiful daughters, Chloé and Kylie, Ilse you were always there to see me progressing in my career from our early high school days right though to the end. You always believed in your man and stood by me on this journey. You and our two girls always had to pay the highest price when I had to sit in front of my computer for hours and hours and also when I traveled to Tilburg. I want to thank you for giving me this opportunity, time and space to follow my dream. I love you and our two children. Last but not least to God who has given me the strength and stood by me.

Deon Meiring, Pretoria, December 2006

"... after climbing a great hill, one only finds that there are many more hills to climb..." Nelson Mandela

Tabel of Contents

Chapter One
General Introduction: History and Societal Context of South African Psychological Assessment

The current thesis addresses bias and equivalence in psychological tests in South Africa. More specifically, the adequacy of the selection procedure used by the South African Police Service (SAPS) is examined. This introduction chapter starts with a brief description of the history and societal context of South African psychological assessment. The second part of the chapter describes issues in cross-cultural assessment and presents a model of bias and equivalence. Finally, a brief description of the current project is given.

Bias and Equivalence of Psychological Measures in South Africa

Psychological testing in South Africa cannot be separated from the country's political, economic, and social history, according to Claassen (1997). Psychological testing developed in several distinct time periods with often different foci. The historical overview in this section mainly pays attention to societal factors that had a bearing on psychological assessment in South Africa. In the beginning of the twentieth century, South Africa was a British colony and psychological testing finds its roots in this colonial heritage. Psychological testing followed the same patterns as in Europe and the United States; the cultural context in South Africa in which psychological tests were first developed was an environment characterized by unequal distribution of resources based on racial categories (Blacks, Coloreds, Indians, and Whites). According to Foxcroft and Roodt (2005), the developments of psychological assessment almost inevitably reflected the racially segregated society in which it evolved.

In the period prior to the new Government of National Unity in 1994, the development of psychological tests was shaped by the Apartheid political dispensation (Foxcroft, 1997). In the early period from 1920 - 1960 psychological measures were either adaptations of overseas measures or they were developed for specific categories of White people. Fick (1929), for instance, developed the Fick Scale which was the South African version of the Stanford-Binet instrument and used it for the testing of White schoolchildren. Fick also applied the instrument to a large sample of Black schoolchildren. He found the mean scores of Black pupils to be much lower than those of the White pupils; he attributed the difference to inferior teaching in Black schools, teaching methods, and Black

children's unfamiliarity with the nature of the test. However, in a follow-up study in 1939, Fick attributed the difference in performance on differences in ability between Blacks and Whites. In his book *African Intelligence,* Biesheuvel (1943) was a strong critic of Fick's views. He questioned the cultural appropriateness of Western-type intelligence test for Blacks and highlighted the influence of different cultural, environmental, and temperamental factors and the effects of malnutrition on intelligence.

In summary, this period between 1920-1960 can be characterized by three main features. Firstly, the focus was on standardizing measures for whites only; secondly, there was a widespread misuse of measures by administering measures standardized for one group to another group without investigating whether or not the measures might be biased and inappropriate for the other group. Thirdly, test results were misused to reach conclusions about differences between groups without considering the impact of (inter alia) cultural, socioeconomic, environmental, and educational factors on test performance.

After 1960 assessment of Black people became more systematic and had a more pragmatic focus on the educability and trainability of black South Africans (Bedell, Van Eeden, & Van Staden, 1999). This change was sparked by rapid developments in the manufacturing and mining industry to select and place Black semi-skilled workers. Sociopolitical developments in the latter half of the 1980s led to the abolition of job reservation and the advent of racial mixed schools (Foxcroft, 1997). According to Claassen (1995), during the period of 1980 to 1994 industry and education authorities began to demand common tests that would not discriminate against any race or culture. Anti-testing lobbyists argued that available tests were biased and led to discriminatory practices and should be banned. In an attempt to address these problems two approaches were followed in the period. Firstly, measures and norms were developed for more than one racial group so that test performance could be interpreted in relation to an appropriate norm group. Secondly, measures were developed and standardized on only White South Africans but also used to assess other groups (Foxcroft & Roodt, 2005).The first thorough South African study of bias only took place in 1986. Owen (1986) investigated test and item bias of the Senior Aptitude Test, Mechanical Insight Test, and the Scholastic Proficiency test among Blacks, Whites, Coloreds, and Indian subjects. He found major differences between the test scores of Blacks and Whites and concluded that understanding and reducing the differential performance of Black and White South Africans would be a major challenge. More bias studies were conducted during this period; Abrahams 1996, Owen (1989a, 1989b), Retief (1992), Taylor and Boeyens (1991), and Taylor and Radford (1986) reported the presence of bias in both ability and personality measures.

The period after South Africa's first democratic election in 1994 up to now saw the application, control, and development of assessment measures becoming a contested terrain (Foxcroft & Roodt, 2005). Nzimande (1995), an influential politician and the leader of the communist party in South Africa expressed himself as follows on the matter:

> The context within which testing is going to take place in South Africa has completely changed. South Africa is shifting from being an Apartheid society to a society that is predominantly concerned with addressing and meeting the basic needs of the majority of the people of the country. One of the most important developments is government's commitment to create a human rights culture. This is captured by the existence of a strong Bill of Rights in the constitution. The implications of a Bill of Rights for psychological testing are far-reaching. Testing in South Africa developed within the context of national, racial and gender oppression. No matter how much psychologists might have thought they were practicing their "science" of testing by observing the ethics of this profession, the fact of the matter is that this was not possible in a society that could be characterized as "unethical". (p. 5)

With the adoption of the new Constitution and the Labour Relations Act in 1996, trade unions and individuals now have the support of legislation that specifically forbids any discriminatory practices in the workplace and includes the protection for applicants as they have all the rights of current employees in this regard. The Employment Equity Act No. 55 of 1998 (section 8), Government Gazette, (1998) refers to psychological tests and assessment and states that:

> Psychological testing and other similar assessments are prohibited unless the test or assessment being used (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group.

According to Foxcroft and Roodt (2005), the Employment Equity Act has major implications for assessment practitioners in South Africa because many of the measures currently in use, whether imported from abroad or developed locally, have not been cross-culturally validated. Abrahams and Mauer (1999) report that despite the election of the new government, the promulgation of the new Labour Relations Act and the Employment Equity Bill, the recommendations of Taylor (1987) have not been heeded. Taylor published a special report focusing on test bias and the roles and responsibilities of test user and test publishers. His main conclusion is that item bias is the most pressing responsibility facing the test constructor in South Africa. Other issues in the domain of comparability and bias should also engage test constructors' attention; for example, more research will have to be done on construct validity across

cultures. The use of a number of tests, which have not been properly validated to be used in selection decisions within a multicultural context, is still rife. It is fair to conclude that the problems in establishing and ensuring equity in cross-cultural assessment have not been solved adequately in modern South Africa, despite the recent legislation and increased attention for the topic.

Cross-Cultural Assessment

Cross-cultural psychology is the systematic study of relationships between the culture context of human development and the behaviors that become established in the repertoire of individuals growing up in a particular culture (Berry, Poortinga, Pandey, Dasen, Saraswathi, Segall, & Kagitcibasi, 1996). The field of cross-cultural psychology is diverse, some psychologists work intensively within one culture, some work comparatively across cultures, and some work with ethnic groups within culturally plural societies, all seeking to provide an understanding of these cultural relationships. Cross-cultural assessment has emerged as a very popular research area and plays an important role in cross-cultural studies, as test scores provide the basis for cross-cultural comparisons, which are the target of the investigation.

Van de Vijver (2002) referred to cross-cultural assessment as all issues arising in the application of psychological instruments, either in a single country in the assessment of migrant groups, or in the assessment of individuals from at least two countries. According to Van de Vijver, it is essential that the tests used have demonstrated their appropriateness in all cultural groups involved.

There are different theoretical perspective employed in the cross-cultural assessment literature. The three dominant perspectives towards assessment are *cross-cultural*, *cultural*, and *indigenous* (Church, 2001). The cross-cultural approach typically involves the following: (a) comparisons of multiple cultures in the search for cultural universals or culture-specific amidst universals; (b) treatment of culture, or quantitative variables related to ecology and culture, as variables outside the individual which can be used to predict behavior; (c) use of traditional and relatively context–free psychometric scales and questionnaires; (d) concern about the cross-cultural equivalence of constructs and measures; and (e)a focus on individual differences. The cultural psychological approach involves the following: (a) a focus on contextual descriptions of psychological phenomenon in one or more cultures, with less emphasis on, or expectations of, culture universals; (b) a theoretical emphasis on the dynamic and mutually constitutive nature of culture and psychological functioning; (c) an emphasis on qualitative, ethnographic, and interpretive research methods; and (d) a de-emphasis on individual differences. Finally, the indigenous approach focuses on the need to formulate theory, constructs, and methods that reflect indigenous cultural context.

As the current project is comparative in nature and addresses the adequacy of instruments in a multilingual and multicultural context, the cross-cultural approach was used. The key concepts of this approach will be discussed.

Bias and Equivalence

From a methodological perspective the most characteristic features of cross-cultural assessment are bias and equivalence. Van de Vijver (2003) indicates that bias and equivalence are often treated as antonyms. Bias is the same as nonequivalence, and equivalence refers to the absence of bias. According to Van de Vijver and Tanzer (1997), bias occurs when score differences in the indicators of a particular construct do not correspond with differences in the underlying trait or ability. Equivalence involves the implications of bias on the scope for comparing scores. So, bias refers to the presence of nuisance factors, which impact on the scores obtained with some instrument, while equivalence is the concept to describe the consequence of the nuisance factors on the comparability of scores across cultures; bias refers thus to unwanted though systematic sources of variation.

Internal Bias

A distinction can be made between two different forms of bias; the first, internal bias, focuses on the relationship between an observed score and a latent trait variable. Internal bias refers to the presence of nuisance factors that play a differential role in different cultures. For example, scores of a questionnaire may be more influenced by social desirability in one culture than in another. Internal bias challenges the validity of comparisons of constructs or scores obtained in different cultural groups. The second form is external bias (also known as predictive bias or differential prediction) and focuses on the relationship between two observed variables – a predictor (e.g., cognitive test or personality measure) and a criterion (e.g., a performance instrument or training performance). If a test shows external bias, the accuracy of statements about which applicants should be accepted and rejected is moderated by culture.

Van de Vijver and Leung (1997) identified three different types of internal bias (see Table 1.1). The *first is called construct bias*; it occurs when the construct measured is not identical across groups or when behaviors that constitute the domain of interest from which items sampled, are not identical across cultures. The second, called *method bias*, is due to various methodological aspects of a study and three types can be discerned, sample bias, instrument bias, and administration bias. The *last type of bias is item bias* (or differential item functioning) and refers to anomalies at item level.

Table 1.1 *Types of Bias and Equivalence*

Type	Definition
Bias	Presence of validity-threatening factors.
Internal bias	Focuses on the relationship between an observed score and a latent trait variable.
Construct bias	The construct measured is not identical across cultural groups.
Method bias	All sources of bias derived from method aspects (e.g., incomparability of samples, instrument inadequacy, and procedural problems).
Item bias	Persons with the same standing on the underlying construct (e.g., they are equally intelligent) but coming from different cultural groups, do not have the same average score on the item.
External bias	Focuses on the relationship between two observed variables, a predictor and a criterion, that differs across cultures.
Equivalence	Comparability of the constructs underlying and measurement scale constituted by the test.
Construct non-equivalence	No comparison possible; comparing "apples and oranges".
Structural equivalence	The test measures the same constructs in all cultural groups.
Measurement unit equivalence	Measurement scales have the same units of measurement and different origin across cultural groups.
Full score equivalence	The same interval or ratio scales applies to all cultural groups

Van de Vijver and Leung (1997) identified four levels of equivalence (see Table 1.1). First, *construct nonequivalence* is a consequence of the presence of construct bias. Second, *structural equivalence* is primary based on similarity in correlations across a variety of cultures, but not necessarily on the same quantitative scale is the second. Third, in the case of *metric or measurement unit equivalence*, the same construct is measured on a scale with identical metrics, but not necessarily with the same scale origin. Fourth, with scalar or *full-score equivalence*, the same construct is measured on an identical interval or ratio

scale. The level of equivalence defines the basis of cross-cultural comparisons and as such qualifies the interpretation of culture differences.

According to Van de Vijver and Leung (1997), equivalence is not an intrinsic property of measurement, but rather dependent on the instrument and culture groups examined. As a consequence, equivalence of measures used for cross-cultural comparisons should be empirically established rather than presumed. This thesis deals with the empirical assessment of cross-cultural bias and equivalence of psychological measures in South Africa.

The Current Project

The current project attempts to add to the body of knowledge in the field of cross-cultural assessment in South Africa. The four studies reported here make use of large samples representing respondents of all cultural groups in South Africa. The data were collected from large applicant pools of applicants who pursued jobs in the South African Police Service over the period of 2000 to 2003. This is one of the first studies to systematically address internal and external bias in high-stake testing across language groups in South Africa. The study also takes cognizance of both the implications of employment equity legislation on police selection and the fact that police officials need to be psychologically healthy as this is a precondition for fulfilling their responsibilities in an adequate manner. Emotional or psychological conditions might unfavorably affect competent performance on the job and even more significantly, endanger the lives of others. Pre-employment psychological screening by SAPS is a presumably effective way to select those applicants who will be successful and competent to become police officials.

South African Police Services Selection Process

The selection process of the SAPS is one of the biggest selection initiatives undertaken by an organization in South Africa. Since the World Cup Soccer will be held in South Africa in 2010, large numbers of applicants have to be selected, hired, and trained in order to have a sufficiently large pool of police employees available within the next few years. For instance, during the next three years, 34,850 recruits will be trained to become police officials. On average, about 30 people apply for one police position (Meiring, 2005); as a consequence, large numbers of applications have to be dealt with in the SAPS selection procedure.

Assessment within the SAPS is done by one of its departments, called Psychological Services. Immediately after democratization in 1994, there was a moratorium on the recruitment and assessment initiatives in the SAPS. The moratorium on applicant assessment was lifted in 1997 with the amalgamation of

the various police agencies into one South African police service. Since 1997, the SAPS has launched recruitment drives every year, culminating in escalating numbers of applications being received which are now up to 3 million applications.

The SAPS are required to have equity plans in place which cater for an ethnically equitable representation in all occupational categories and levels in the workforce. Equity targets for entry-level constables are set prior to the selection process and are aligned with the affirmative action plan of the SAPS. Post allocation of entry-level positions is based on three criteria, namely demographic composition of the country, composition of the population from where entry-level applicants are recruited, and organizational needs. In most cases, 80% of the posts are allocated to Blacks followed by Coloreds, Asians, and Whites. Potential police applicants are informed about police jobs through local newspapers. These job announcements contain information concerning the necessary requirements (e.g., age, completed matric as minimum educational qualification, driver's license, no criminal record, and physical and medical fitness) and the selection procedure. Once the SAPS have received application forms, the recruitment offices screen them for compliance with the minimum requirements. Eligible applicants are then invited to attend a one-day selection session. The selection takes place in all nine provinces in South Africa (i.e., Gauteng, North West Province, Mpumalanga, Northern Province, Northern Cape, Free State, Western Cape, Eastern Cape, and KwaZulu Natal). Provinces are further divided up into smaller areas where recruitment offices are situated. Applicants are tested at the closest recruitment office to their place of residence. Every applicant goes through a standardized selection procedure.

During the one-day selection, a selection battery is administered that consists of cognitive measures (e.g., reading and comprehension and a spelling test) and a personality test (e.g., 15FQ+ personality test). The cognitive ability and the personality tests are paper-and-pencil tests and are provided in English. Each applicant's finger prints are also taken and are checked against a criminal record database for any offences. The last activity in the selection process is a physical screening test (a job-related 500-meter obstacle course). Applicants need to complete the course within a certain time limit. After the one-day selection, the electronic answer sheets are sent to Psychological Services in Pretoria (Head Office), where the answer sheets are electronically scanned. A "multiple hurdle approach" is followed with the test battery where applicants have to attain a certain score level on the cognitive measures. In the next phase the applicant needs to score within certain boundaries on the personality profile. A cut off score for the physical screening test are also set. A selection decision making model is utilized to generate a long list for each of the culture groups that is rank ordered.

A further shortlist is compiled after applicants have gone through a medical assessment on a different time.

After the selection procedure, applicants who have been selected on the basis of their ranking start with the job training. They receive a theoretical training of six months and a practical weapon training of four months. After the completion of this training program, they are stationed at a police station and receive field training for another six months. After a two-year trainee period they are hired, become police officers, and receive full benefits such as a medical scheme and allowances. The entire training program is conducted in English. The selection process attempts to be efficient, speedy, and objective. The objectivity of the procedure is mainly achieved by not relying on interviewers or other assessors. Secondly, the procedure is objective in that scores are statistically combined into a final ranking expressing each applicant's chance to be selected.

Research Design

I am interested in the question of whether the test battery used by the SAPS complies with the Employment Equity Act. In terms of cross-cultural assessment, this compliance refers to the absence of internal and external bias. More specifically, the current project addressed four questions, namely:

1. To what extent are the current instruments being used by the South African Police Services unbiased and equivalent?
2. Is it possible to adapt existing tests in order to make them free of bias (or at least substantially reduce the bias) for all South African groups?
3. Does our test battery comply with standards regarding internal and external bias?
4. Can we develop a new instrument that is free of bias (or at least shows as little bias as possible) so as to enable a comparison of South African results to international studies and findings?

This research project included four empirical studies (each described in a separate chapter of this thesis). The first two studies mainly address internal bias and examine structural equivalence and the influence of test adaptations on equivalence. The third study addresses external bias. The last study examines to what extent we can build an appropriate instrument by combining existing measures. The four studies together provide a test of how adequate existing instruments are with regards to bias and equivalence within the South African context.

The thesis ends with a concluding chapter based on discussion of the main findings.

References

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F., & Mauer, K. F. (1999). The comparability of the constructs of the 16PF in the South African context. *SA Journal of Industrial Psychology, 25,* 53-59.

Bedell, B., Van Eeden, R., & Van Staden, F. (1999). Culture as moderator variable in psychological test performance: Issues and trends in South Africa. *SA Journal of Industrial Psychology, 25*(3), 1-7.

Berry, J. W., Poortinga, Y. H., Pandey, J., Dasen, P. R., Saraswathi, T. S., Segall, H. S., & Kagitcibasi, C. (1996*). Handbook of cross-cultural psychology* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Biesheuvel, S. (1943). *African intelligence.* Johannesburg: South African Institute of Race Relations.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69,* 979-1006.

Claassen, N. C. W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47*, 297-307.

Claassen, N. C. W. (1995, October). *Cross-cultural assessment in the human sciences.* Paper presented at a work session on the meaningful use of psychological and educational tests. Pretoria. Human Sciences Research Council.

Fick, M. I. (1929). Intelligence test results of poor white, native (Zulu), coloured and Indian school children and the educational and social implications. *South African Journal of Science, 26*, 904-920.

Foxcroft, C. D., & Roodt, G. (2005). *An introduction to psychological assessment in the South African context* (2nd ed.). Oxford, UK: Oxford University Press.

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13,* 229-235.

Meiring, D. (2005, April). *Police selection in South Africa.* Presented at a meeting of the Erasmus University Rotterdam, The Netherlands.

Nzimande, B. (1995, June). *Culture fair testing. To test or not to test?* Paper presented at the Psychometrics Congress. Pretoria, South Africa.

Owen, K. (1986). *Test and item bias: Administration of the Senior Aptitude Test, Mechanical Insight Test and the Scholastic Proficiency Battery to White,*

Indian, Black and Coloured Technikon students. Pretoria: Human Sciences Research Council.

Owen, K. (1989a). The suitability of Ravens's Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13,*149-159.

Owen, K. (1989b). *Test and item bias: The suitability of the Junior Aptitude Test as a common test battery of White, Indian and Black pupils in standard seven.* Pretoria: Human Sciences Research Council.

Retief, A. (1992). The cross-cultural utility of the SAPQ: Bias or fruitful differences? *South African Journal of Psychology, 17,* 202-207.

Taylor, T. R., & Boeyens, J. C. A. (1991). *A comparison of black and white responses to the South African Personality Questionnaire.* Pretoria: Human Sciences Research Council.

Taylor, T. R. (1987). *Test bias: The roles and responsibilities of test user and test publisher.* Johannesburg. National Institute for Personnel Research.

Taylor, J. M., & Radford, E. J. (1986). Psychometric testing as an unfair labour practice. *South African Journal of Psychology, 16,* 79-86.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Van de Vijver, F. J. R. (2002). Cross-cultural assessment: Value for money?. *Applied Psychology: An international Review, 51,* 545-566.

Van de Vijver, F. J. R. (2003). Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143-155). New York: Wiley.

Chapter Two
Construct, Item, and Method Bias of Cognitive and Personality Tests in South Africa*

Abstract

Bias was studied for two cognitive tests and a personality test at three levels: the construct underlying the test ("construct bias"), method-related aspects such as response sets ("method bias"), and the items ("item bias"). The sample consisted of 13 681 participants who had applied for entry-level jobs in the South African Police Service. The cognitive instruments produced very good construct equivalence and low item bias. However, various scales of the personality questionnaire revealed construct bias in various ethnic groups. The item bias in the personality scales was low. Method bias did not have any impact on the (small) size of the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably in the black groups.

*Meiring, D., Van de Vijver, A.J.R, Rothmann, S., & Barrick, M.R. (2005). Construct, item, and method bias of cognitive and personality measures in South Africa. *South African Journal of Industrial Psychology, 31(1), 1-8.*

Introduction

Psychological testing in South Africa cannot be investigated in isolation without taking the country's political, economic, and social history into account (Claassen, 1997). Psychometric testing in South Africa has mainly followed international trends and at the beginning of the 1900s tests were imported from abroad and applied in all sectors of the community (Foxcroft, 1997). Cross-cultural issues emerged in the 1920s, and in the 1940s and 1950s psychological testing focused on the educability and trainability of black South Africans. In the 1980s certain aspects of fairness, bias, and discriminatory practices received more attention in line with international developments. Separate psychological tests were initially developed for the Afrikaans and English-speaking groups (Claassen, 1997). At a later stage bilingual tests were constructed for English and Afrikaans speakers and separate tests were constructed for speakers of African languages.

Since the first democratic elections, held in 1994, the country has had a new constitution and stronger demands for the cultural appropriateness of psychological tests culminated in the promulgation of the Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998, p. 9), which stipulates the following: "Psychological testing and other similar assessments are prohibited unless the test or assessment being used (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group."

The onus of proof has shifted to psychologists using these instruments, who now have to indicate that they adhere to the regulations of the Employment Equity Act. Given the transformation of the South African society, the integration of schools, universities, the work place, and society in general since 1994, there is an urgent need for measuring instruments that meet the Employment Equity Act requirements and can be used for all the cultural and language groups in South Africa.

The current study examines the extent to which the most important tests in the assessment procedure to recruit new police officials for the South African Police Services (SAPS) – two cognitive tests (a Reading and Comprehension Test and a Spelling Test) and a personality questionnaire (15FQ+) meet the criteria imposed by the Employment Equity Act by examining bias in the instruments employed.

Bias and equivalence

Bias and equivalence are pivotal concepts in the application of psychological tests in a multicultural society such as South Africa. According to Van de Vijver and Tanzer (1997), bias occurs when score differences in the indicators of a particular construct do not correspond with differences in the underlying trait or ability. Equivalence on the other hand refers to score comparability, namely the measurement level at which scores obtained for different cultures can be compared. Consequently, bias refers to the influence of nuisance factors (unwanted but systematic sources of variation) in cross-cultural score comparisons whereas equivalence is the consequence of the nuisance factors concerning the comparability of scores across cultures. Van de Vijver and Tanzer (1997) note that bias has to do with the characteristics of an instrument in a (specific) cross-cultural comparison rather than with its intrinsic properties. The question as to whether an instrument is biased cannot be answered in general terms, but can be addressed when an instrument is biased in a specific comparison.

Van de Vijver and Leung (1997a, 1997b) propose a taxonomy of bias consisting of three types, namely construct bias, method bias and item bias. Construct bias occurs when the construct measured is not identical across cultures or when behaviours that characterise the construct are not identical across cultures. This type of bias can stem from several sources; for example the definition of a construct may show an incomplete overlap across cultures. Method bias refers to problems caused by the manner in which a study is conducted (method-related issues). Three types of method bias can be distinguished (Van de Vijver, 2002). First, incomparability of samples on factors other than the target variables can lead to method bias (sample bias). Second, method bias also refers to problems arising from instrument characteristics (instrument bias). Third, method bias arises from administration problems (administration bias). Item bias (also referred to as differential item functioning) refers to the situation in which the (psychological) meaning of one or more items is not identical across cultures and relates to anomalies at the item level, such as poor translation or inapplicability of an item to a specific culture.

Van de Vijver and Tanzer (1997) consider bias as an indication of a source of systematic cross-cultural differences that need to be studied. Bias analysis can offer important clues concerning the causes of cross-cultural differences and can thus be regarded as a phenomenon that requires further explanation. According to Van de Vijver and Leung (1997a, b), equivalence refers to the implications of bias with regard to the comparability of constructs and test

scores. Van de Vijver and Tanzer (1997) treat equivalence from a measurement perspective and make a hierarchical distinction between three types of equivalence. The first level is called construct equivalence. This means that the same construct is measured across all cultural groups studied, irrespective of whether or not the measurement of the construct is based on identical instruments across cultures. It implies the universal validity of the underlying psychological construct. The second level of equivalence is called metric or measurement unit equivalence and is obtained when two metric measures have the same measurement unit but different origins. In the case of measurement unit equivalence no direct score comparisons can be made across cultural groups unless the size of the offset (i.e., the difference in scale origin) is known. The highest level of equivalence is scalar equivalence or full-scale equivalence and this is obtained if two metric measures have the same measurement unit and the same origin.

Bias and equivalence in cognitive and personality tests in South Africa

Cognitive tests. Cross-cultural comparison of cognitive test scores is not new in South Africa (Irvine, 1969). Biesheuvel's (1943, 1954) early work in South Africa focuses on the empirical investigation of potential bias problems associated with cross-cultural assessment. Biesheuvel emphasised the importance of home environment, schooling, nutrition, and other factors in cognitive test performance in a multicultural society. Schepers (1974) reported that urban subjects, when compared with rural examinees, have a slightly greater differentiated intellect, with education playing the biggest role in the differentiation process. Freeman (1984) reported that the cognitive skills needed to deal with the Raven Progressive Matrices are better developed in an urbanised population than in a rural one. Verster and Prinsloo (1988) compared the results of IQ points of different generations and found decreasing differences between the English speaking and Afrikaans speaking adults. Claassen (1997) reported that between 1954 and 1984 the mean difference between English-speakers and Afrikaans-speakers was reduced from ten IQ points to five IQ points. Socioeconomic and educational circumstances change from one generation to another and have an impact on cognitive test scores. This phenomenon contributes to method bias.

In South Africa few studies focused on the construct equivalence of cognitive measures across cultures. Most studies that were carried out concerned comparisons between English speakers and Afrikaans speakers. A high degree of structural equivalence was reported in these studies (Cudeck & Claassen, 1983; Verster, 1974; Vorster 1978). Between 1960 and 1984 it was not necessary for psychology to look at the issue of construct equivalence since tests were

developed independently for each of the race groups and no cross-cultural comparisons were made (Claassen, 1997; Owen, 1992). In the 1980s there was growing interest in comparing cultural groups with regard to existing cognitive tests. Claassen (1993) applied the New South African Group Test (NSGT) to Blacks, Coloureds, Indians, and Whites in order to assess the cross-cultural suitability of the test. All the respondents wrote the test in English. The verbal part of the test was problematic for the Black group since English was not their mother tongue. Large mean differences were reported for the cultural groups and the structural equivalence was found to be poor. Owen (1986) investigated structural equivalence and item bias by applying three cognitive tests (Senior Aptitude Test, Mechanical Insight Test and Scholastic Proficiency Battery) to Black, Coloured, Indian, and White students. He reported structural equivalence across these cultural groups and item bias analyses supported the suitability of the measures for all groups. Owen (1989) also examined the structural equivalence and item bias of the Junior Aptitude Test for White, Indian and Black pupils in Standard 7. For the Black pupils the structural equivalence was problematic. Many items in the case of the Indian and Black groups were biased. Results pointed to the strong influence of education and understanding of the English language on structural equivalence and of item bias on cognitive tests.

Personality questionnaires. Cross-cultural personality research has focused extensively on the universality of the five-factor model (FFM) (Cheung et al., 2001; McCrae & Allik, 2002; Paunonen, Zeidner, Enggvik, Oosterveld, & Maliphant 2000; Roland, Parker, & Strumf, 1998) and Eysenck's three-factor model (Barrett, Petrides, Eysenck, & Eysenck, 1998). In South Africa a few studies have been conducted, investigating the FFM across cultural groups. Heuchert, Parker, Strumf, and Myburg (2000) applied the NEO-Personality Inventory-Revised (NEO-PI-R) to college students. The authors found a clear five-factor solution for both Black and White students. An unpublished thesis (Horn, 2000) examined a Xhosa translation of the NEO-PI-R. Horn reported that translation was difficult and that various items could not be translated into Xhosa because of its restricted vocabulary. Taylor (2000) carried out a construct comparability study of the NEO-PI-R for Black and White employees in a work setting. The NEO-PI-R did not work as well for Blacks as it did for Whites. In particular the openness factor could not be extracted in the Black sample. Other studies in South Africa made use of the South African Personality Questionnaire (SAPQ) and the 16 PF (South African 1992 version). There was little support for construct equivalence across the different cultural groups in South Africa. Individuals whose first language was not English experienced problems with the

questionnaire, especially because some of the items were difficult to understand. Researchers concluded that these tests were not suitable for use in a multicultural society like South Africa (Abrahams, 1996, 2002; Abrahams & Mauer, 1999a, 1999b; Meiring, 2000; Spence, 1982; Tact 1999; Taylor & Boeyens, 1991).

In summary, cognitive and personality cross-cultural studies had seldom been carried out in South Africa before the 1980s. In line with international trends there has been increasing interest in the topic during the last few decades. Structural equivalence and item bias of cognitive tests were studied while in the case of personality tests the focus was mainly on structural equivalence. These studies mainly adopted the designs and statistical procedures found in the Anglo-Saxon literature (Berry et al., 2000). Studies in South Africa reported race, education, language, and understanding of English as the main reasons impacting on construct and item comparability of cognitive and personality tests. There is a need to continue to research the issues of bias in a contemporary South Africa.

Research aims

The first research aim of this study was to examine bias at the level of constructs (structural equivalence) and items (item bias) in two cognitive tests and a personality test that were administered to select entry-level police officials for the South African Police Service (SAPS). In addition, method bias was studied by examining the influence of cognition and social desirability on the 15FQ+.

Method

Participants

The sample consisted of 13,681 participants throughout South Africa who applied for entry-level police jobs in the SAPS. Applicants came from all nine provinces. The sample consisted of Blacks (n = 11,626), Whites (n = 570), Indians (n = 662) and Coloureds (n = 812). Ninety percent (n = 11,317) were male and ten percent (n = 2,353) were female. The Black group consisted of the following nine ethnicities: Ndebele (n = 259), Sepedi (n = 1,777), SeSotho (n = 1,285), Setswana (n = 2,009), Swati (n = 294), Tsonga (n = 922), Venda (n = 978), Xhosa (n = 1,725), and Zulu (n = 2,404). The mean age of the sample group was 25 years (SD = 2.8). The entry-level requirement for the police is Grade 12, 69% of the sample had a Grade 12 qualification, 13% had a degree or diploma, and 18% had a post-graduate qualification.

Instruments

The test battery consisted of a cognitive section, which included an English reading and comprehension test, an English spelling test and the 15FQ+ Questionnaire.

The reading and comprehension test consisted of four paragraphs that were selected from the basic training modules (Module 1: the Bill of Rights on Police Power, Community Policing; Module 2: Non-Verbal Communication; Module 5: Mental Disorders). Five questions were asked in respect of each paragraph. The test requires the applicant to read the paragraphs and comprehend the material in order to answer the questions. The test consists of 20 items and each item has four response alternatives. A time limit of 20 minutes was allowed for the completion of the test. The spelling test was also developed for the SAPS. Training instructors at the training college were asked to generate a pool of police-relevant words (such as *rape* and *homicide*) which students find difficult to spell when they start their basic training. A pool of words was generated and a spelling test consisting of 40 items was developed. An item consisted of four different spellings of a single word. Applicants had to select the correctly spelled word. A time limit of 12 minutes was given for the completion of the test. The reliability of reading and comprehension and spelling test (internal consistency; Cronbach's alpha) for the different language groups is reported in Table 2.1. The mean alpha coefficients of the two tests are 0.84 (spelling test) and 0.64 (reading and comprehension), respectively. All these values are acceptable ($\alpha > 0.60$, Clark & Watson, 1995), and thus indicate an acceptable internal consistency

Table 2.1 *Values of Cronbach's Alpha across Cultural Groups per Test/Scale*

	Cultural Group											
Test/ Scale	Xhosa	Zulu	Ndebele	Sepedi	SeSotho	Setswana	Swati	Tsonga	Venda	Indian	Coloured	White
Cognitive												
Reading and Comprehension Test	.623	.634	.647	.601	.564	.633	.607	.618	.586	.697	.685	.764
Spelling Test	.841	.840	.827	.854	.838	.842	.834	.854	.823	.837	.816	.849
Personality Scales												
Cool Reserved – Outgoing	.429	.445	.396	.510	.510	.457	.510	.527	.474	.643	.559	.629
Intellectance	.551	.529	.452	.551	.576	.518	.465	.583	.501	.670	.639	.615
Affected by Feelings - Emotionally Stable	.590	.596	.552	.581	.638	.652	.603	.567	.627	.750	.730	.753
Accommodating – Dominant	.286	.383	.364	.326	.356	.377	.349	.328	.230	.655	.587	.680
Sober Serious – Enthusiastic	.546	.603	.477	.569	.611	.606	.621	.568	.500	.688	.700	.758
Expedient – Conscientious	.472	.501	.468	.485	.465	.460	.428	.450	.537	.683	.537	.624
Retiring – Socially Bold	.638	.629	.602	.599	.629	.637	.609	.553	.518	.818	.746	.826
Tough Minded – Tender Minded	.384	.345	.406	.354	.403	.448	.388	.348	.279	.712	.628	.755
Trusting – Suspicious	.353	.364	.354	.351	.392	.385	.415	.364	.356	.682	.607	.700
Practical – Abstract	.088	.138	.245	.091	.154	.114	.182	.0006	.118	.447	.388	.461
Forthright – Discreet	.421	.453	.530	.502	.480	.479	.420	.491	.421	.667	.564	.698
Self-assured – Apprehensive	.355	.404	.460	.434	.453	.444	.460	.426	.420	.267	.378	.283
Conventional – Radical	.231	.157	.268	.199	.163	.151	.003	.160	.005	.478	.346	.532
Group – Orientated – Self-Sufficient	.507	.560	.544	.524	.549	.552	.519	.496	.421	.702	.665	.760
Undisciplined – Self-Disciplined	.375	.400	.401	.436	.362	.315	.392	.391	.383	.382	.384	.405
Relaxed – Tense Driven	.429	.455	.396	.506	.510	.457	.510	.527	.474	.643	.559	.629

The 15FQ+ is a normative, trichotomous response, personality test that has been developed by Psytech International as an update of the original 15FQ (Tyler, 2002). Both versions of the 15FQ were designed for use in industrial and organizational settings. The original version of this assessment was first published in 1991 as an alternative to the 16PF series of tests. The original 15FQ was designed to assess 15 of the 16 personality dimensions that were first identified by Cattell and his colleagues in 1946. The 15FQ+ is a complete revision of the original 15FQ, with the authors developing and fielding a completely new item set for the 15FQ+. The authors' stated aim was to produce a relatively short, yet robust measure of Cattell's primary personality factors (Tyler, 2002). It has been known for some time that reasoning ability (or intelligence) cannot be reliably measured by reasoning items included in untimed personality tests, as is the case with Cattell's Factor B. For this reason Factor B was excluded from the 15FQ. However, in the case of the 15FQ+, the authors decided to deal with this problem by redefining Factor B as a "metacognitive personality variable" called intellectance. Validity and reliability have been determined for the 15FQ+ (Tyler, 2002). For this study the reliabilities (Cronbach's alpha) for the different language groups are reported in Table 2.1. The internal consistencies for some of the factors were very low, notably in the Black language groups. There is a serious problem with the internal consistencies of the following factors: Practical – Abstract (mean alpha = 0.20) and Conventional – Radical (0.22) across all groups. These low values seriously challenge the suitability of the 15FQ+ in this multicultural setting.

Procedure

Applicants were tested in groups of 100 during April 2000. A standardised procedure was followed by previously trained personnel of the Psychological Services of the SAPS in order to apply the test battery. The test session lasted for three hours and also contained a break of 15 minutes. Computer-readable answer sheets were utilised for all the tests.

Statistical Analysis

Construct bias and item bias were addressed in two series of analyses for both the cognitive and personality tests. The first involved scale-level analyses and examined the similarity of the factors underlying the cognitive and personality tests, whereas the second addressed bias at item level of the instruments. Method bias in the personality scales was examined by looking at the influence of cognition and social desirability on the personality scores.

Scale-level analysis (construct bias). A two-step procedure was used to examine construct bias which is based on exploratory factor analysis. In the first step the covariance matrices of all the cultural groups were combined (weighted by sample size) in order to create a single, pooled data matrix (cf. Muthén, 1991, 1994). Factors derived from this pooled covariance matrix define the global solution, with which the factors obtained in the separate cultural groups were compared (after target rotation to the pooled solution). The agreement was evaluated by means of a factor congruence coefficient, Tucker's phi (Chan, Ho, Leung, Cha & Yung, 1999; Van de Vijver & Leung, 1997a, 1997b). Values above 0.90 are taken to point to essential agreement and values above 0.95 to very high agreement. High agreement implies that the factor loadings of the lower and higher level are equal up to a multiplying constant. (The latter is needed to accommodate possible differences in the eigenvalues of factors for the language groups).

Item level analysis (item bias analysis). Item bias analysis was undertaken by using two different procedures. Logistic regression was used for the cognitive instruments (yielding dichotomous scores) and analysis of variance (ANOVA) was used for the personality test (yielding interval-level scores). Both kinds of analyses are based on the same conceptualization of item bias. The assumption is that an item is unbiased if persons from different cultures with an equal standing on the theoretical construct underlying the instrument have the same expected score on the item (Van de Vijver & Leung, 1997a, 1997b).

Logistic regression is a general procedure of analysing differential item functioning (DIF) as it can detect both uniform and non-uniform bias (Mellenbergh, 1982; Van de Vijver & Leung, 1997a, 1997b) in dichotomous items and thus provide a model-based approach for studying DIF (Rogers & Swaminathan, 1990, 1993). The total test score (a proxy for ability level) and culture are the independent variables, while the item score is the dependent variable. The presence of a significant main effect of score level is usually taken as an indication of uniform bias. An item is taken to show non-uniform bias if the interaction between level and culture is significant. In the present study the sample size was large so that conventional tests of significance could not be used. The procedure that was used for the cognitive tests computed the effect size for the items, where the difference between the Nagelkerke R^2 of the first step (in which score level was the sole predictor) and second step (in which culture, dummy coded was added as a predictor) provides an estimate of the effect size of culture (uniform bias). In the third step the interaction of culture and

score level is added; the difference between the second and the third estimates the impact of the interaction (non-uniform bias).

In the analysis of variance of the personality items the item score was the dependent variable, while culture and score levels were the independent variables. Analogous to the previous analysis, a significant main effect of the culture group was taken to point to uniform bias, and a significant interaction of score level and culture interaction pointed to non-uniform bias.

Finally, the influence of the presence of biased items on the size of cross-cultural differences was examined. This was done by comparing the cross-cultural differences in the original 15FQ+ questionnaire with those in the 15FQ+ questionnaire from which presumably biased items had been removed.

Method bias analysis. Method bias was studied in respect of the personality questionnaire. From the literature it could be concluded that knowledge of the English language could be an important moderator of responses to the 15FQ+. Similarly, differences in response styles across cultural groups could also be expected to exert some influence. In order to examine their impact, a multivariate analysis of covariance was carried out. Cultural group (12 levels) was the independent variable; the dependent variables were the scale scores of the 15FQ+ while cognitive ability (as a proxy for English language proficiency, which was the testing language) and social desirability were the covariates.

Results

Scale-Level Structural Equivalence

Cognitive tests. Based on a scree test, both cognitive tests showed a unifactorial solution in the pooled data. Table 2.2 shows the agreement of the factor derived from the pooled data with the factor in the 12 language groups for both cognitive tests. Values of Tucker's phi higher than 0.90 were found in the two tests for all the language groups. This provided a strong indication of the structural equivalence of the cognitive factors underlying the performance of all the different groups distinguished.

Table 2.2. *Agreement of the Reading and Comprehension Test in the Pooled Solution with the Black Group Divided into Nine Language Groups and the Three other Race Groups (Tucker's phi)*

	Test	
Language Group	Reading and Comprehension	Spelling
Xhosa	0.992	0.998
Zulu	0.975	0.990
Ndebele	0.957	0.907
Sepedi	0.990	0.995
SeSotho	0.990	0.994
SetSwana	0.995	0.989
Swati	0.975	0.975
Tsonga	0.991	0.991
Venda	0.984	0.984
Indian	0.974	0.965
Coloured	0.992	0.975
White	0.966	0.975

Personality. Scree tests of the factor analyses of the separate scales suggested the extraction of a single factor in each analysis. The agreement of the factors of the 15FQ+ in the pooled solution with factors in the 12 language groups is indicated in Table 2.3. Various entries in the table showed values well below the threshold level of 0.90. More specifically, a column comparison revealed that for four of the groups there were problems with the structural equivalence of the constructs (Ndebele 50%, Whites 44%, Indians 31%, and Coloureds 25% of the factors). A row comparison showed that in particular two scales, Conventional – Radical and Relaxed – Tense Driven, did not show structural equivalence across six of the groups. Only three scales showed equivalence across all of the language groups: Accommodating – Dominant, Retired – Socially Bold, Group Orientated–Self-Sufficient.

Table 2.3 *Agreement of the 16 Factors in the Pooled Solution with the Black Group Divided into Nine Sub-Language Groups and the Three other Race Groups*

Factor	Xhosa	Zulu	Ndebele	Sepedi	SeSotho	Setswana	Swati	Tsonga	Venda	Indian	Coloured	White
Cool Reserved – Outgoing	.989	.992	.882	.998	.995	.993	.980	.996	.992	.986	.989	.974
Intellectance	.997	.998	.942	.989	.997	.991	.959	.995	.992	.952	.987	.891
Affected by Feelings - Emotionally Stable	.997	.998	.951	.991	.995	.998	.994	.975	.988	.980	.990	.971
Accommodating – Dominant	.969	.985	.838	.948	.963	.992	.924	.983	.736	.972	.962	.908
Sober Serious – Enthusiastic	.989	.989	.948	.992	.994	.995	.957	.983	.979	.950	.968	.952
Expedient – Conscientious	.983	.993	.859	.987	.976	.991	.915	.970	.988	.980	.974	.956
Retiring – Socially Bold	.995	.996	.927	.991	.998	.994	.976	.984	.984	.995	.991	.990
Tough Minded – Tender Minded	.983	.958	.956	.948	.976	.994	.937	.907	.947	.814	.851	.780
Trusting – Suspicious	.985	.993	.895	.989	.992	.989	.957	.971	.988	.819	.737	.582
Practical – Abstract	.995	.997	.943	.994	.994	.991	.944	.992	.981	.945	.910	.806
Forthright – Discreet	.966	.980	.862	.982	.993	.988	.948	.991	.959	.962	.937	.953
Self-assured – Apprehensive	.993	.982	.975	.984	.987	.987	.936	.987	.976	.893	.961	.877
Conventional – Radical	.853	.988	.705	.913	.966	.970	.877	.940	.962	.352	.441	.400
Group - Orientated – Self-Sufficient	.988	.996	.965	.989	.990	.993	.979	.959	.986	.980	.993	.971
Undisciplined – Self-Disciplined	.985	.994	.615	.986	.987	.978	.940	.980	.954	.935	.939	.928
Relaxed – Tense Driven	.901	.969	.761	.938	.851	.930	.916	.929	.895	.821	.825	.847

Item-Level Analyses

Cognitive tests. It is clear from Table 2.4 and 2.5 that when bias is evaluated in terms of significance, many items revealed significant bias (reading and comprehension 50%, spelling test 68%). Cohen's (1988) criteria according to which the lower threshold for medium-size effects is 0.06 was applied to further examine the size of the item bias (this size was chosen as it can be considered to be significantly large to be practically important). It was found that for the reading and comprehension test only one item out of 20 showed non-uniform bias and for the spelling test item one item out of 40 items showed uniform bias. It can be concluded that many items show statistical bias but the bias effect is so slight as to be negligible from a practical perspective.

Table 2.4 *Items with Bias of Effect Size and Significance for the Reading and Comprehension for the Different Language Groups*

Item	Uniform bias	Non-uniform bias
1	0.010	0.003*
2	0.007*	0.001
3	0.004*	0.002*
4	0.001	0.002
5	0.004*	0.002
6	0.007*	0.005*
7	0.003	0.001
8	0.012	0.002*
9	0.003	0.001
10	0.002*	0.001
11	0.004*	0.001*
12	0.006	0.001
13	0.005*	0.003*
14	0.003*	0.001
15	0.002	0.002
16	0.002*	0.003*
17	0.005	0.001
18	0.004*	0.002*
19	0.004*	0.005*
20	0.002	0.001

*$p < 0.05$ (item shows significant (non-)uniform bias if followed by an asterisk)

Table 2.5 *Items with Bias of Effect Size and Significance of the Spelling Test for the Different Language Groups*

Item	Uniform bias	Non-uniform bias	Item	Uniform bias	Non-uniform bias
1	0.031*	0.002	21	0.012*	0.005*
2	0.012*	0.002*	22	0.005*	0.008*
3	0.005*	0.003*	23	0.002	0.001
4	0.012	0.002	24	0.003	0.002
5	0.009*	0.002*	25	0.015*	0.003*
6	0.030*	0.002*	26	0.019*	0.011*
7	0.007*	0.005*	27	0.038*	0.003*
8	0.007*	0.001	28	0.007	0.001
9	0.003*	0.002*	29	0.004	0.003*
10	0.006*	0.004*	30	0.005	0.002
11	0.063*	0.006*	31	0.012*	0.002*
12	0.022	0.004*	32	0.006*	0.002
13	0.007	0.002	33	0.008	0.002
14	0.048*	0.002*	34	0.002	0.002*
15	0.006*	0.003*	35	0.006*	0.005*
16	0.002*	0.003*	36	0.058*	0.001*
17	0.013*	0.001*	37	0.005*	0.003*
18	0.022	0.002	38	0.005*	0.003*
19	0.008*	0.002*	39	0.015	0.002
20	0.010	0.001	40	0.026*	0.001*

$p <$ 0.05 (item shows significant (non-)uniform bias if followed by an asterisk)

Personality. In analyses of variance of the item scores of the 15FQ+ we found that many items showed a significant main effect of culture (uniform bias) or interaction of culture and score level (non-uniform bias). Out of the 200 items, 72 turned out to be biased (36%), which is a large proportion. However, only one item showed a medium effect size. It can be concluded that item bias is not a major disturbance in the 15FQ+ in these language groups.

Influence of bias on size of cross-cultural differences. In order to inspect the impact of item bias on cross-cultural differences in the personality scales, the size of these differences was computed before and after the elimination of biased items. An item was taken to be biased if it had an eta square value of at least 0.02 for the uniform or non-uniform bias component. This low value was used because of the overall low level of the effect sizes. One-way analyses of variance were carried out with language group as independent variable and scale scores (sum scores on the items pertaining to the scale) as dependent variables. In a second step the procedure was repeated, but now all biased items were excluded from the computation of scale scores. The extent of the cross-cultural differences was evaluated as the effect size (eta square) of the culture component. The mean effect size was 0.027 before the removal of biased items and 0.028 after bias removal. It could be concluded that the correction for biased items did not affect the size of the cross-cultural differences observed.

Method Bias in the Personality Questionnaire

In order to evaluate the impact of method bias the effects of cognitive/language ability and social desirability were scrutinized in an analysis of covariance. The size of the cross-cultural differences was computed before and after correction for the covariates (ability and social desirability). The main effect of the cross-cultural difference was 0.026 before correction and 0.025 after correction for covariates. Clearly the results of covariate analysis revealed that cognitive ability and social desirability scores did not have any impact on the size of the cross-cultural differences of the personality questionnaire.

Discussion

This study was the first South African study in which different types of bias were studied: bias at the level of constructs, items, and the method of administration. The sample consisted of 13,681 participants throughout South Africa who had applied for entry-level police jobs in the SAPS. The sample was split into 12 different language groups. A police-specific cognitive test containing

subtests of reading/comprehension and spelling test, and a personality questionnaire, the 15FQ+, were administered in this study.

Both cognitive measures showed low levels of construct bias; both revealed factorial invariance in all the language groups. Item bias analyses showed several items revealing significant bias. Instead of the significance of item bias indicators, their effect size was used as the criterion to evaluate the presence of item bias (this was done because of the large sample size). If the presence of a medium or large effect size for the indicators of uniform or non-uniform bias is taken as the criterion of item bias, almost no items showed significant bias. It seems fair to conclude that the extent of item bias is not very consequential in the cognitive measures.

The examination of the construct bias of the personality measures showed less favourable results. Structural equivalence was particularly problematic for the two factor scales (Conventional – Radical, Relaxed – Tense Driven) in four language groups (Whites, Coloureds, Indians, and Ndebele). The item bias analyses did not point to major problems at item level in any personality scale. Not surprisingly, the removal of the biased items did not affect the size of the cross-cultural differences observed. An analysis of the influence of cognitive ability (as a proxy for English language proficiency) and social desirability (as a measure of response style) revealed that the extent of the cross-cultural differences between the language groups was not influenced by these factors, thereby suggesting that the influence of these sources of method bias could be safely ignored in the current data.

The Anglo-Saxon literature, often reporting studies done in the U.S.A., provides support for the structural equivalence of most cognitive tests (Berry et al., 2002). However, for personality questionnaires the equivalence picture is not so clear (Ellis, 1995). In this study high levels of structural equivalence were found for the cognitive tests but in the case of personality test structural equivalence across the different language groups was problematic for the 15FQ+. The current results are fairly consistent with the mainstream literature. The findings with respect to item bias in the cognitive tests are also in line with the mainstream literature (Berk, 1982; Holland & Wainer, 1993): many items were found to be biased, but the size of the bias is small as is its impact on the size of intergroup differences. Similarly, the personality questionnaires showed many biased items, but their size was small and their impact on observed scores obtained in the various language groups very limited. One of the reasons for the small size of the bias may be the educational entry-level requirement, which apparently reduced the educational heterogeneity of the sample considerably. As

a consequence, the results may not be generalizable to a broader, more unselected sample of the South African population. Even though the bias was small, the current findings underscore Church's (2001) conclusion that a major challenge for cross-cultural personality studies is that equivalence of constructs and measures will rarely, if ever, be fully met.

A serious problem concerns the low internal consistencies (more in the Black than the other groups). The reliability values of various personality scales are so low that they cannot be adequately used for individual assessment and selection purposes.

The nature of the construct bias of some of the personality scales was further explored, using an expert group, consisting of Black SAPS psychologists and two African language experts. They were asked to identify aspects of the personality measures that might be a threat to the structural equivalence. Several aspects of the questionnaire were mentioned, such as the level of the words being used and the understanding of the context and interrelationship of words could be problematic, especially for Black groups (e.g., analytical, intellectually, conventional, gullible, genuinely, temperamental, smashing). The use of double meanings in items could cause confusion. The use of idiomatical expression raised concerns (e.g., "both feet firmly on the ground", "head in the clouds"). Qualifying words such as "rarely", "generally", "less", and "on occasion" could also be problematic. Finally, it was pointed out that some of the constructs could be more culture specific. Looking at the history of South Africa for example the construct of Conventional – Radical will have a stronger political connotation for the Black respondents than for other groups. Relaxed – Tense, African respondents can be seen as more relaxed people than others.

Prinsloo and Ebersohn (2002) argue that different response rates to personality items could reflect real differences in underlying traits. In the case of personality traits, which often comprise of highly socialised constructs, it is reasonable to expect that various additional sources contribute to intergroup differences. What role do education and the understanding of English play in the construct bias of the 15FQ+. Abrahams and Mauer (1999b) qualitatively examined the impact of home language on the responses to the items of the 16PF. They concluded that the understanding of items and concepts in English was problematic, especially for Black groups. Prinsloo and Ebersohn (2002) proposed that by testing respondents' English proficiency can help to assess its impact on performance in personality measurement.

Does the present study answer the question of whether the test battery being used by the SAPS to select entry-level applicants can stand the scrutiny of

the Employment Equity Act 55 of 1998 and its subsections (Government Gazette, 1998)? The cognitive tests did not show much bias, whereas some personality scales were problematic. Moreover, various personality scales showed unacceptably low internal consistencies. Consequently, the results of the cognitive tests are encouraging, whereas an uninformed application of the personality scales could be problematic. In addition to problems with the structural equivalence, there is the additional problem of low internal consistencies - one more issue that challenges the use of the scales in selection. If personality constructs can be identified that are important to a police official, a selective strategy can be followed and factors that did not show structural equivalence and factors with unacceptably low internal consistencies can be avoided.

The current study did not address all aspects of test usage. More specifically, the predictive validity and predictive bias of the tests were not considered. Even an unbiased instrument may not work equally well for different language groups. The current study did not address the question whether the cognitive and personality scales can predict future training and job performance in a fair way for all language groups. A final verdict on the cross-cultural suitability of the current test battery can only be given when data on the predictive bias are available. Although the jury is still out, the prospects for the personality instrument are dim because of its low reliability in notably the Black groups.

References

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo & I. Ebersohn. *South African Journal of Psychology, 32,* 58-61.

Abrahams, F., & Mauer, K. F. (1999a). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology, 25,* 53-59.

Abrahams, F., & Mauer, K. F. (1999b). Qualitative and statistical impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South African context. *South African Journal of Psychology, 29,* 76-86.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25,* 805-819.

Berk, R. A. (1982). *Handbook of methods for detecting item bias.* Baltimore, MD: Johns Hopkins University Press.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology. Research and applications* (2nd ed.). Cambridge: Cambridge University Press.

Biesheuvel, S. (1943). *African intelligence.* Johannesburg: South African Institute of Race Relations.

Biesheuvel, S. (1954). The measurement of occupational aptitudes in a multi-racial society. *Occupational Psychology, 52,* 289-196.

Chan, W., Ho, R. M., Leung, K., Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378-402.

Cheung, F. M., Leung, K., Zhang, J. X., Sun. H. F., Gan, Y. Q., Song, W. Z., & Xie, D. (2001). Indigenous Chinese personality constructs. *Journal of Cross-Cultural Psychology, 32,* 407-433.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69,* 979-1006.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309-219.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Claassen, N. C. W. (1993). *Verslag oor die funksionering van die NSAG intermedier G in verskillende bevolkingsgroepe.* Pretoria: Raad vir Geesteswetenskaplike Navorsing.

Claassen, N. C. W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47,* 297-307.

Cudeck, R., & Claassen, N. C. W. (1983). Structural equivalence of intelligence tests for two language groups. *South African Journal of Psychology,13,* 1-5.

Ellis, B. B. (1995, August). *Examination of the measurement equivalence of a Spanish version of the 16 PF using Item Response Theory.* Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Freeman, M. C. (1984). *The effect of cultural variables on the Goodenough-Harris Drawing Test and the Standard Progressive Matrices.* Unpublished master's dissertation, University of the Witwatersrand, Johannesburg, South Africa.

Foxcroft, C. D. (1997) Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment, 13,* 229-235.

Government Gazette, Republic of South Africa, Vol. 400, no. 19370. Cape Town, 19 October 1998.

Heuchert, J. W. P., Parker, W. D. Strumf, H., & Myburg, C. P. H. (2000). The five-factor model for African college students. *American Behavioral Scientist. 44,* 112-125.

Holburn, P. T. (1992). *Differential item functioning in mental alertness test.* Unpublished master's dissertation, University of South Africa, Pretoria, South Africa.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Horn, B. S. (2000). *A Xhosa translation of the revised NEO Personality Inventory.* Unpublished master's dissertation, University of Port Elizabeth, Port Elizabeth, South Africa.

Irvine, S. H. (1969). Factors analysis of Africa abilities and attainments: Constructs across cultures. *Psychological Bulletin, 71,* 20-32.

McCrae, R. R., & Allik, J. (Eds.). (2002). *The five–factor model of personality across cultures.* New York: Kluwer Academic / Plenum Publishers.

Meiring, D. (2000, June). *Revisiting the cross-cultural comparability of the 16 Personality Factor Inventory (16PF) in the South African context.* Paper presented at the Industrial Psychology Conference (incorporating the Psychometrics Conference), Pretoria, South Africa.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7* 105-118.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376-398.

Owen, K. (1986). *Test and item bias: Administration of the Senior AttitudeTest, Mechanical Insight test and the Scholastic Proficiency Battery to White, Indian, Black and Coloured Technikon students.* Pretoria: Human Sciences Research Council.

Owen, K. (1989). *Test and item bias: The suitability of the Junior Aptitude Test as a common test battery of White, Indian and Black pupils in standard seven.* Pretoria: Human Sciences Research Council.

Owen, K. (1992). *Test-item bias: Methods, findings and recommendations.* Pretoria: Human Sciences Research Council Group: Education

Paunonen, S. V., Zeidner, M., Enggvik, H., Oosterveld, P., & Maliphant, R. (2000). The nonverbal assessment of personality in five cultures. *Journal of Cross-Cultural Psychology, 31*, 220-239.

Prinsloo, C. H., & Ebersohn, I. (2002). Fair usage of the 16PF in personality assessment in South Africa: A response to Abrahams and Mauer with special reference to issues of research methodology. *South African Journal of Psychology, 32* , 48-57.

Rogers, H. J., & Swaminathan, H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal for Educational Measurement, 27*, 361-370.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17,* 105-117.

Roland, J. P., Parker, W. D., & Strumf, H. (1998). A psychometric examination of French translation of the NEO-PI–R and NEO-FFI. *Journal of Personality Assessment, 71*, 269-29.

Schepers, J. M. (1974). Critical issues which have to be resolved in the construction of tests for developing groups. *Humanitas RSA, 2*, 395-406.

Spence, B. A. (1982). *A psychological investigation into the characteristics of black guidance teachers.* Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

Tact, H. (1999). *The cross-cultural validity and compatibility of the Sixteen Personality Factor Questionnaire.* Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

Taylor, I. A. (2000). *The construct comparability of the NEO PI-R Questionnaire for Black and White employees.* Unpublished doctoral thesis, University of the Free State, Bloemfontein, South Africa.

Taylor, T. R., & Boeyens, J. C. A. (1991). *A comparison of black and white responses to the South African Personality Questionnaire.* Pretoria: Human Sciences Research Council.

Tyler, G. (2002). *A review of the 15FQ+ Personality Questionnaire.* Pulloxhill, UK: Psychometrics Limited.

Van de Vijver, F. J. R. (2002). Cross-cultural assessment: Value for money? *Applied Psychology: An International Review, 51,* 545-566.

Van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed.). (pp. 257-300). Boston, MA: Allyn & Bacon.

Van de Vijver, F. J. R., & Leung, K. (1997b). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological Issues. *Journal of Personality, 69,* 1006-1030.

Van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Verster, J. M. (1974). A study of intellectual structure in two groups of South African scientists. *Psychologia Africana, 15,*169-190.

Verster, J. M., & Prinsloo, R. J. (1988). The diminishing test performance gap between English speakers and Afrikaans speakers in South Africa. In S. H. Irvine, S. H. Berry, J. W. (Eds.) *Human abilities in cultural context* (pp. 534-560). Cambridge: Cambridge University Press.

Vorster, J. F. (1978). The invariance of the factor structure of the intermediate and senior forms of the NSAGT. *Humanitas*, 3, 409-417.

Chapter Three
BIAS IN AN ADAPTED VERSION OF THE 15FQ+ QUESTIONNAIRE IN SOUTH AFRICA*

Abstract

A previous study examined bias at construct and item level in a personality inventory (15FQ+) among a large sample of over 15,000 Whites, Indians, Coloureds, and nine Black groups who had applied for entry-level police jobs at the South African Police Service (Meiring, Van de Vijver, Rothmann, & Barrick, 2005). An instrument with adaptations based on the findings of this study was administered to a sample comparable in size and ethnic composition to the sample of the previous study. With the exception of the Relaxed – Tense Driven Scale, which remained problematic, the adapted version produced less construct and item bias than the original version. In addition, the adapted instrument revealed only marginal increases in internal consistencies, notably the Black groups' consistency levels remained low. It was concluded that bias in the adapted 15FQ+ was reduced but the low internal consistencies continued to limit the usefulness of the inventory.

*Meiring, D., Van De Vijver, F. J. R. , Rothmann, S. (2006). *Bias in an Adapted Version of the 15FQ+questionnaire in South Africa. South African. Journal for Psychology, 36, 340-356.*

Introduction

Cross-cultural assessment in South Africa has been placed on the agenda with the promulgation of the Employment Equity Act 55 of 1998, Section 8 (Government Gazette, 1998). According to this Act,

> "Psychological testing and other similar assessments are prohibited unless the test or assessment being used - (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group".

The Employment Equity Act imposes very strict criteria on South African psychologists. The onus of proof shifts to professional test users, as they have to indicate that their instruments adhere to the regulations of the Employment Equity Act and can be applied in a multicultural society. Multicultural personality research in South Africa is virtually nonexistent (Abrahams, 1996, 2002; Abrahams & Mauer, 1999a, b; Meiring, 2000; Spence, 1982; Tact, 1999; Taylor, 2000; Taylor & Boeyens, 1991; Wallice & Birt, 2003). Research by Abrahams (1996) on the cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF) has received most of the attention to date in South Africa (Prinsloo & Ebersohn, 2002). Little support was found for the equivalence of the 16PF across the different cultural groups in South Africa. Individuals whose first language was not English experienced problems with the comprehensibility of the items. It was concluded that this test was not suitable for use in South Africa.

Meiring, Van de Vijver, Rothmann, and Barrick (2005) recently investigated the adequacy of cognitive tests and the Fifteen Factor Questionnaire (15FQ+), a personality measure, in a group of police applicants from all major South African ethnic groups. Construct, method, and item bias were examined. Construct bias refers to the question whether the same underlying constructs are measured in each ethnic group; method bias is a generic term for instrument-related and person-related factors that can systematically affect the size of cross-cultural score differences, such as differential social desirability; item bias refers to the presence of items that do not measure the same in each cultural group (cf. Van de Vijver & Leung, 1997a, b). Several scales of the personality questionnaire revealed construct bias in various ethnic groups. The item bias in the personality scales was low, while method bias did not have any impact on the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably in the Black groups. It was concluded that the 15FQ+ was not suitable as an instrument in the South African

multicultural context because of the low internal consistencies of some scales and the lack of construct equivalence.

The first study suggested various ways to adapt the 15FQ+ so as to increase its cross-cultural suitability. The current study examines bias in an adapted version of the 15FQ+ that was part of a test battery used to recruit new police officials for the South African Police Services (SAPS) in 2001 and examines whether these tests meet the criteria imposed by the Employment Equity Act by examining bias in the adapted instrument.

Test Adaptation to Reduce Bias

According to Hambleton (1994), cross-cultural research using solid psychometric tests is on the increase and with this growth, the need has arisen to adapt (commonly called "translate") instruments to be used in multiple cultures and languages. The work of developing test adaptation guidelines has been carried out by an international committee of psychologists who are affiliated to various international psychology associations working in the field of cross-cultural psychology (such as the International Association for Cross-Cultural Psychology). This committee has developed and validated guidelines for adapting psychological instruments and established score equivalence across language and/or culture groups (Hambleton, 1994, 2001). According to the committee, the term "adaptation" rather than "translation" was preferred as translation is just a part of the more encompassing adaptation process. Van de Vijver (2003) indicates that there has been an important trend in the design of cross-cultural studies in the last decade. Where in the past the translation of instruments was often seen as primarily a linguistic issue, there is now a growing awareness that translating an instrument has to be seen in a wider context and requires an expertise in the language and culture of the specific target group as well as in item writing (Hambleton, 1994, 2001; Hambleton, Merenda, & Spielberger, 2004; Van de Vijver & Hambleton, 1996). It may be noted that the committee worked from the implicit assumption of a multilingual test administration. However, their work is also relevant for studies in which a single language version is used among multiple ethnic groups in which the testing language is not the mother tongue.

The new development amounts to an increasing awareness of the impact of bias on cross-cultural comparisons. The new trend of adaptation of tests refers to a team effort (the so-called "committee approach"; see Harkness, 2003; Van de Vijver & Leung, 1997a, b) where experts from different disciplines (linguistic, cultural, and psychometric) join forces. Adaptation of the instrument in this new

approach is part of a whole chain of activities, aimed at maximizing the quality of the use of the instrument in a cross-cultural setting.

Adapted tests are those in which some component has been deliberately altered independent of unavoidable translation change. These adaptations may be substantive, relate to question design, or consist of slight wording modifications. Regardless of the form of change, the aim of adaptation is to render questions culturally or linguistically appropriate in a cross-national context. According to Van de Vijver (2003), adaptation is seen to offer several advantages, such as their sensitivity to the cultural context in which they are used, their ease and relatively low price to produce, and their flexibility to deal with major sources of bias. A potential problem with adaptation is that the equivalence of modified items must be tested and demonstrated. A further drawback is that adapted items provide limited scope for statistical analysis, since only few statistical tools can be used as these have to be able to accommodate partly dissimilar items.

Research aims

Adapting the original version of the 15FQ+ in South Africa is important in view of the serious problems associated with the construct bias and low internal consistency of the instrument reported in our earlier study (Meiring et al., 2005). Studies in South Africa report education, language and proficiency in English as the main factors impacting on construct and item comparability of personality tests. It is important to evaluate the adequacy of adapted tests and to determine to what extent these are successful in dealing with the bias sources identified in the original version. The first research aim of this study was to examine the construct and item bias of the adapted version of the 15FQ+, as well as its internal consistency. As method bias did not affect the scores on the original instrument, the current authors decided not to analyse its impact in the adapted version. In addition, the results of the original 15FQ+ (Meiring et al., 2005) were compared with the results of the adapted version of the 15FQ+. The impact of reading comprehension (part of the cognitive test battery) as a moderator on the adapted version of the 15FQ+ was also investigated.

Method

Participants

The sample included 16,339 participants across South Africa who had applied to become a police official in the SAPS. The sample consisted of Blacks (n = 14,415), Whites (n = 579), Asians (n = 378) and Coloureds (n = 1,065) (44 cases missing). Eighty percent (n = 13,091) were male and 19% (n = 3,015) were

female. The Black group consisted of the following groupings: Ndebele (n = 393), Sepedi (n = 2 974), SeSotho (n = 1,272), Setswana (n = 1,996), Seswati (n = 411), Tsonga (n = 1,237), Venda (n = 1,098), Xhosa (n = 1,947), and Zulu (n = 2,817). We will refer to our subsamples, based on a combination of race and ethnic/linguistic groups, as cultural groups in the remainder of the article. The mean age of the sample group was 25 years (SD = 2.70). The entry level qualification for the police is Grade 12; 78% of the sample group had this qualification, 18% had a degree or diploma and 1% had a postgraduate qualification (3% missing).

Instruments

The test battery consisted of a cognitive section, which included an English reading and comprehension test and a personality test, consisting of the adapted 15FQ+ Questionnaire. The cognitive tests (reading and comprehension test) were developed specifically for the SAPS. The reading and comprehension test consisted of four paragraphs that were selected from the basic training modules (Module 1: the Bill of Rights on Police Power, Community Policing; Module 2: Non-Verbal Communication; Module 5: Mental Disorder). Five questions were asked about each paragraph, making a total of 20 items. The test required the applicant to read the paragraphs and comprehend the material in order to answer the questions. Each item has four response alternatives. A time limit of 20 minutes was allowed for the completion of the test. The reliability of the reading and comprehension test (internal consistency; Cronbach's alpha) for the different language groups is reported in Table 3.1.

Table 3.1 *Values of Cronbach's Alpha and Full Range across Cultural Groups per Test/Scale*

Test/ Scale	Xhosa	Zulu	Ndebele	Sepedi	Sesotho	Setswana	Swati	Tsonga	Venda	Asian	Coloured	White
Cognitive												
Reading and Comprehension Test	0.63	0.67	0.56	0.61	0.60	0.62	0.62	0.58	0.59	0.69	0.68	0.77
Personality Scales												
Cool Reserved – Outgoing	0.47	0.55	0.55	0.50	0.54	0.51	0.53	0.50	0.54	0.66	0.58	0.62
Intellectance	0.60	0.59	0.55	0.53	0.57	0.61	0.59	0.55	0.49	0.71	0.58	0.63
Affected by Feelings - Emotionally Stable	0.61	0.59	0.54	0.55	0.64	0.64	0.58	0.56	0.54	0.74	0.68	0.70
Accommodating – Dominant	0.29	0.38	0.35	0.31	0.31	0.35	0.17	0.28	0.24	0.71	0.57	0.65
Sober serious - Enthusiastic	0.55	0.63	0.56	0.57	0.61	0.62	0.56	0.55	0.54	0.69	0.66	0.71
Expedient – Conscientious	0.48	0.53	0.49	0.49	0.52	0.50	0.49	0.50	0.50	0.67	0.61	0.62
Retiring - Socially Bold	0.63	0.65	0.63	0.58	0.63	0.64	0.51	0.53	0.50	0.82	0.72	0.80
Tough Minded – Tender Minded	0.43	0.48	0.52	0.42	0.45	0.48	0.36	0.44	0.39	0.74	0.67	0.74
Trusting – Suspicious	0.48	0.48	0.53	0.46	0.48	0.50	0.51	0.48	0.44	0.77	0.61	0.48
Practical – Abstract	0.14	0.15	0.07	0.18	0.17	0.12	0.06	0.12	0.16	0.44	0.31	0.49
Forthright – Discreet	0.51	0.50	0.44	0.50	0.58	0.50	0.53	0.51	0.55	0.68	0.63	0.71
Self-assured – Apprehensive	0.27	0.29	0.31	0.31	0.33	0.33	0.38	0.33	0.29	0.69	0.56	0.71
Conventional – Radical	0.37	0.38	0.31	0.32	0.32	0.41	0.16	0.28	0.34	0.59	0.50	0.54
Group-orientated – Self-Sufficient	0.59	0.62	0.61	0.53	0.57	0.59	0.51	0.51	0.50	0.67	0.62	0.76
Undisciplined – Self-Disciplined	0.38	0.35	0.31	0.40	0.38	0.38	0.35	0.39	0.44	0.40	0.32	0.35
Relaxed – Tense Driven	0.41	0.39	0.40	0.36	0.33	0.38	0.31	0.41	0.39	0.78	0.64	0.78

The 15FQ+ is a personality test that has been developed by Psytech International as an update of the original 15FQ (Tyler, 2002). The original 15FQ was designed to assess 15 of the 16 personality dimensions that were first identified by Cattell and his colleagues in 1946. The new test, called 15FQ+, is a complete revision of the original 15FQ. The authors developed and fielded a completely new item set for the 15FQ+. The authors' aim was to produce a relatively short, yet robust measure of Cattell's primary personality factors (Tyler, 2002). It has been known for some time that reasoning ability (or intelligence) cannot be adequately measured by reasoning items included in untimed personality tests, as is the case with Cattell's Factor B. For this reason Factor B was excluded from the 15FQ. However, in the case of the 15FQ+, the authors decided to deal with this problem by redefining Factor B as a "metacognitive personality variable" called intellectance. Validity and reliability have been determined for the 15FQ+. Good reliability coefficients of between 0.60 to 0.85 were found for student and professionals in the UK. Tyler, 2002 reported similar reliabilities (with a mean of 0.75) were reported for South African professional and management development candidates.

The adaptation of the 15FQ+ was a team effort, following an approach outlined by Harkness (2003) and Van de Vijver and Leung (1997a, b). The different experts consisted of the CEO of Psytech South Africa (15FQ+ test distributor in South Africa) who is a registered research, counselling and industrial psychologist, the first author who is a senior industrial psychologist in the Psychological Services of the SAPS, and a linguistic and culture expert, who has ample experience with speakers of English for whom this language is not the mother tongue at various levels of proficiency. The adaptations made by the committee focused primarily on the item question design and wording modifications to the 15FQ+. Items and words were adapted to be more culturally and linguistically appropriate for the South African context to reduce bias. A total of 44% (85 items) of the original 15FQ+ items were changed; in 38 items, one or two words were changed and in 47 other items the complete item stem was reworded. Three constructs (Practical – Abstract, Conventional – Radical, Relaxed – Tense Driven) received most of the attention; 21 items of these scales (out of a total of 36 items) were rephrased and 11 items were reworded slightly. Examples of changed items and words in items were the following: "Everybody has their price" was changed to "Anybody can be bribed in some way or another"; and "I find myself deeply engrossed in thought" was changed to "I find myself deeply lost in ideas".

Procedure

Applicants were tested in groups of 100 in June 2001. The instruments were administered using a standardized procedure by previously trained personnel of Psychological Services of the SAPS. The test session lasted 3 hours and a break of 15 minutes was allowed. Computer-readable answer sheets were employed for all tests.

Statistical analysis

Internal consistencies of the adapted 15FQ+ for the different culture groups were calculated. As reading and comprehension skills were considered to be important moderators, reliabilities were recalculated restricting the analysis to applicants with a score above a stanine of 8 on the reading and comprehension test.

Construct bias and item bias were addressed in two series of analyses for the adapted 15FQ+ questionnaire. The first involved scale-level analyses and examined the similarity of the factors underlying the personality measures, while the second addressed bias at the item level of the instruments.

Scale-level analysis *(construct bias)* can be investigated with several techniques, such as factor analysis, cluster analysis, and multidimensional scaling or other dimensionality-reducing techniques (Van de Vijver & Leung, 1997a, b). The basic idea behind the application of these techniques is to obtain a structure in each culture, which can then be compared across all cultures involved. Factor analysis is the most frequently employed technique for studying construct equivalence. In the current study both exploratory and confirmatory models could have been used. A choice for confirmatory factor analysis may seem obvious because there is information about the composition of the instruments (on the basis of previous studies). However, the current authors used exploratory factor analysis for a pragmatic reason. They found serious problems with the use of confirmatory models in studies involving large samples in many cultural groups. The main problem in the application of confirmatory models is their fit to the data, which is almost always very bad. It is usually not clear whether the reasons for the poor fit are serious and should lead to a reformulation of the model or whether these are trivial and do not challenge the underlying model.

A two-step procedure, based on exploratory factor analysis, was used to examine construct equivalence. In the first step the covariance matrices of all the cultural groups were combined (weighted by sample size) so as to make a single, pooled data matrix (cf. Muthén, 1991, 1994; Van de Vijver & Poortinga, 2002). Factors derived from this pooled covariance matrix define the global solution with

which the factors obtained in the separate cultural groups were compared (after target rotation to the pooled solution). The agreement was evaluated by a factor congruence coefficient, Tucker's phi (Chan, Ho, Leung, Cha, & Yung, 1999; Van de Vijver & Leung, 1997a, b). Values above 0.90 are taken to point to essential agreement and values above 0.95 to very good agreement. A high agreement implies that the factor loadings of the lower and higher level are equal up to a multiplying constant (the latter is needed to accommodate possible differences in eigenvalues of factors for the language groups).

Item level analysis *(item bias analysis)* was performed by using analysis of variance (ANOVA) for the adapted 15FQ+ questionnaire (yielding interval-level scores). The assumption is that an item is unbiased if persons from different cultures with an equal standing on the construct underlying the instrument have the same expected score on the item (Van de Vijver & Leung, 1997a, b). There are several statistical techniques available for analysing item bias in three-point Likert scales. Probably the most important are item response theory, in which the three categories are treated as ordered responses (e.g., Maydeu-Olivares, Drasgow, & Mead, 1994; Thissen, 1984) and analysis of variance. Item response theory models have the advantage of having detailed and statistically rigorous fit indices; in addition, it is possible to check whether the distance between disagree and undecided is the same as the distance between agree and undecided. The use of analysis of variance has various advantages, such as computational simplicity, robustness, and the possibility to study both uniform and non-uniform bias (Mellenbergh, 1982). In particular, the latter is relevant here. Therefore, analysis of variance was used in the present study. The item score was the dependent variable, while culture and score levels were the independent variables. A significant main effect of culture group was taken to point to uniform bias and a significant interaction of score level and interaction pointed to non-uniform bias.

Results

Internal consistency

Particulars of Cronbach's alpha can be found in Table 3.1. The reliabilities for the reading and comprehension test remained largely at the same level as reported in the Meiring et al. (2005) study. In the adapted 15FQ+, a slight improvement of about 0.02 in the reliabilities was found for the different groups compared to the original version (see Table 3.2). It can be concluded that the internal consistencies for some of the factors still remained very low, notably in the African language groups, and that the increase in internal consistencies was

too low to have any practical salience. This seriously challenges the suitability of the adapted 15FQ+ questionnaire in a multicultural setting. In a subsequent analysis, the authors addressed the influence of reading comprehension as a moderator on the internal consistencies of the 15FQ+. They computed the alpha values for the whole group, and for those respondents with a stanine score of 8 and higher on the reading and comprehension test. The cultural groups were split up into Asian, Black, Coloured and White respondents. As can be seen in Table 3.3, the test adaptations did not have a major impact on the alpha values; neither the values of the total group nor the group with the high stanine scores were much higher after adaptation.

Scale-level structural equivalence

The agreement of the (unifactorial) scales of the 15FQ+ adapted version in the pooled solution with factors in the 12 language groups is indicated in Table 3.4. Various entries in the table showed that 14 factors had values well below the threshold level of 0.90. More specifically, a column-wise comparison revealed that in each group there were one or two constructs with a high construct bias (and hence, low structural equivalence) across the groups. However, these constructs were different across the language groups. A row-wise comparison showed that two scales in particular, Undisciplined – Self-discipline and Relaxed – Tense, were problematic. When comparing the values of the original version and the adapted version of the 15FQ+ three conclusions emerged (see Table 3.5). First, a higher level of overall structural equivalence was reported for the adapted version, with the average of Tucker's phi increasing by about 0.02. Secondly, the Ndebele group, in which 50% of the scales of the previous version showed poor structural equivalence, revealed much higher values of Tucker's phi (the increase was on average 0.09). For the Indian, Coloured, and White groups there were fewer scales with low values of Tucker's phi. Thirdly, one factor, Relaxed – Tense Driven, still showed a poor structural equivalence after adaptation (in 50% of the groups the value of Tucker's phi was less than 0.90).

Table 3.2 *Increase of Internal Consistencies after Adaptation per Cultural Group (Difference of Internal Consistency Adapted Scale and Original Scale)*

Test/ Scale	Xhosa	Zulu	Ndebele	Sepedi	SeSotho	SetSwana	Swati	Tsonga	Venda	Asian	Coloured	White	Mean
Cool Reserved - Outgoing	0.04	0.11	0.15	-0.01	0.03	0.05	0.02	-0.03	0.07	0.02	0.02	-0.01	0.04
Intellectance	0.05	0.06	0.10	-0.03	-0.01	0.09	0.12	-0.03	-0.02	0.04	-0.06	0.02	0.03
Affected by Feelings – Emotionally Stable	0.02	-0.01	-0.01	-0.03	0.00	-0.01	-0.02	0.00	-0.09	-0.01	-0.05	-0.05	-0.02
Accommodating – Dominant	0.00	-0.01	-0.02	-0.01	-0.04	-0.02	-0.18	-0.05	0.01	0.05	-0.01	-0.03	-0.03
Sober serious - Enthusiastic	0.00	0.03	0.08	0.00	0.00	0.01	-0.06	-0.02	0.04	0.00	-0.04	-0.05	0.00
Expedient - Conscientious	0.01	0.02	0.02	0.00	0.05	0.04	0.06	0.05	-0.04	-0.01	0.07	-0.01	0.02
Retiring - Socially Bold	0.00	0.02	0.03	-0.02	0.00	0.00	-0.10	-0.02	-0.02	0.00	-0.03	-0.03	-0.01
Tough Minded - Tender Minded	0.05	0.13	0.11	0.07	0.05	0.03	-0.03	0.09	0.11	0.03	0.04	-0.01	0.06
Trusting – Suspicious	0.12	0.12	0.18	0.11	0.09	0.11	0.10	0.12	0.08	0.08	0.00	-0.22	0.07
Practical – Abstract	0.05	0.01	-0.18	0.08	0.02	0.00	-0.12	0.12	0.04	-0.01	-0.07	0.03	0.00
Forthright – Discreet	0.09	0.04	-0.09	0.00	0.10	0.02	0.11	0.02	0.13	0.01	0.06	0.01	0.04
Self-assured – Apprehensive	-0.08	-0.11	-0.15	-0.13	-0.13	-0.11	-0.08	-0.09	-0.13	0.42	0.18	0.43	0.00
Conventional – Radical	0.14	0.23	0.04	0.12	0.16	0.26	0.16	0.12	0.33	0.11	0.15	0.00	0.15
Group - Orientated – Self-Sufficient	0.09	0.06	0.06	0.00	0.02	0.04	-0.01	0.02	0.08	-0.03	-0.05	0.00	0.02
Undisciplined – Self-Disciplined	0.00	-0.05	-0.09	-0.03	0.01	0.07	-0.04	-0.01	0.05	0.02	-0.06	-0.06	-0.02
Relaxed – Tense Driven	-0.02	-0.06	0.01	-0.14	-0.18	-0.08	-0.20	-0.12	-0.08	0.14	0.08	0.15	-0.04
Mean	0.03	0.04	0.02	0.00	0.01	0.03	-0.02	0.01	0.04	0.05	0.01	0.01	0.02

Table 3.3 Increase of Cronbach's Alpha after the Adaptation and Influence of Reading and Comprehension (RC) as a Moderator (per Cultural Group per 15FQ+ Scale)[a]

Test/ Scale	Black		Asian		Coloured		White	
	Total[b]	Stan > 8[b]	Total	Stan > 8	Total	Stan > 8	Total	Stan > 8
Cool Reserved - Outgoing	0.10	-0.02	0.02	-0.01	0.02	0.10	-0.01	-0.01
Intellectance	0.04	-0.01	0.04	0.09	-0.06	-0.09	0.02	-0.04
Affected by Feelings – Emotionally Stable	0.04	-0.03	-0.01	0.03	-0.05	-0.06	-0.05	-0.11
Accommodating – Dominant	-0.13	-0.05	0.05	0.06	-0.01	0.09	-0.03	-0.04
Sober serious – Enthusiastic	-0.01	-0.01	0.00	0.03	-0.04	-0.05	-0.05	-0.02
Expedient - Conscientious	0.02	-0.05	-0.01	-0.04	0.07	0.15	-0.01	-0.10
Retiring - Socially Bold	-0.01	0.00	0.00	0.03	-0.03	-0.04	-0.03	-0.02
Tough Minded – Tender Minded	0.07	0.04	0.03	-0.02	0.04	0.01	-0.01	-0.04
Trusting – Suspicious	0.11	0.04	0.08	0.06	0.00	-0.09	-0.22	-0.05
Practical – Abstract	0.06	0.07	-0.01	-0.08	-0.07	-0.04	0.03	0.05
Forthright – Discreet	0.05	0.03	0.01	0.02	0.06	0.11	0.01	0.01
Self-assured – Apprehensive	-0.11	-0.04	0.42	0.02	0.18	0.03	0.43	0.03
Conventional – Radical	0.17	0.18	0.11	0.01	0.15	0.06	0.00	-0.04
Group – Orientated – Self-Sufficient	0.04	0.07	-0.03	0.03	-0.05	0.01	0.00	0.06
Undisciplined - Self-Disciplined	-0.01	-0.04	0.02	-0.12	-0.06	-0.20	-0.06	0.00
Relaxed - Tense Driven	0.02	0.07	0.14	0.04	0.08	0.05	0.15	0.02
Mean	0.03	0.02	0.10	0.01	0.02	-0.01	-0.02	-0.02

[a]The reading and comprehension test was the same as on the previous version so that no change in alpha can be computed. [b]The label "Total" refers to the total sample, while "Stan > 8" refers to the group with the reading comprehension scores in stanine 8 and higher.

Table 3.4 *Agreement of the 16 factors in the pooled solution with the black group divided into nine sub language groups and the three other culture groups*

Test/ Scale	Xhosa	Zulu	Ndebele	Sepedi	Sesotho	Setswana	Swati	Tsonga	Venda	Asian	Coloured	White
Cool Reserved – Outgoing	1.00	1.00	0.99	0.99	1.00	1.00	0.98	0.99	0.99	0.97	0.99	0.98
Intellectance	1.00	1.00	0.99	1.00	1.00	1.00	0.98	0.98	0.99	0.98	1.00	0.96
Affected by Feelings - Emotionally Stable	1.00	1.00	0.99	0.99	0.99	1.00	0.99	0.99	0.97	0.96	0.99	0.96
Accommodating – Dominant	0.96	0.99	0.94	0.98	0.97	0.98	0.88	0.96	0.91	0.94	0.96	0.91
Sober serious – Enthusiastic	0.99	1.00	0.99	1.00	1.00	1.00	0.94	1.00	0.98	0.97	0.97	0.96
Expedient – Conscientious	0.99	0.99	0.97	1.00	0.99	0.99	0.96	0.98	0.98	0.97	0.98	0.97
Retiring - Socially Bold	1.00	1.00	0.99	0.99	1.00	1.00	0.96	0.98	0.96	0.99	0.99	0.99
Tough Minded - Tender Minded	0.99	1.00	0.99	0.97	0.99	0.99	0.94	0.97	0.95	0.94	0.96	0.90
Trusting – Suspicious	0.99	1.00	0.98	0.99	0.97	0.99	0.99	0.98	0.98	0.96	0.90	0.91
Practical – Abstract	1.00	1.00	0.99	0.99	0.99	1.00	0.97	0.99	0.99	0.94	0.94	0.88
Forthright – Discreet	0.99	0.99	0.96	0.99	0.99	1.00	0.96	0.97	0.97	0.98	0.97	0.98
Self-assured – Apprehensive	0.99	0.99	0.98	0.99	0.99	1.00	0.99	0.99	0.99	0.91	0.96	0.85
Conventional – Radical	0.99	1.00	0.94	0.99	0.99	1.00	0.87	0.98	0.99	0.95	0.98	0.94
Group-Orientated – Self-Sufficient	1.00	0.99	0.98	0.99	0.99	0.99	0.98	0.99	0.97	0.98	0.99	0.98
Undisciplined - Self-Disciplined	0.98	0.99	0.85	0.99	0.99	0.99	0.91	0.99	0.97	0.88	0.80	0.92
Relaxed - Tense Driven	0.61	0.99	0.97	0.87	0.90	0.95	0.97	0.75	0.96	0.75	0.85	0.76

Table 3.5 *Increase of Tucker's Phi (factorial agreement) after the adaptation (per cultural group per 15FQ+ scale)*

Test/ Scale	Xhosa	Zulu	Ndebele	Sepedi	SeSotho	SetSwana	Swati	Tsonga	Venda	Indian	Coloured	White	Average
Cool Reserved – Outgoing	0.01	0.01	0.11	-0.01	0.01	0.01	0.00	-0.01	0.00	-0.02	0.00	0.01	0.01
Intellectance	0.00	0.00	0.05	0.01	0.00	0.01	0.02	-0.02	0.00	0.03	0.01	0.07	0.02
Affected by Feelings - Emotionally Stable	0.00	0.00	0.04	0.00	-0.01	0.00	0.00	0.02	-0.02	-0.02	0.00	-0.01	0.00
Accommodating – Dominant	-0.01	0.01	0.10	0.03	0.01	-0.01	-0.04	-0.02	0.17	-0.03	0.00	0.00	0.02
Sober Serious – Enthusiastic	0.00	0.01	0.04	0.01	0.01	0.01	-0.02	0.02	0.00	0.02	0.00	0.01	0.01
Expedient – Conscientious	0.01	0.00	0.11	0.01	0.01	0.00	0.04	0.01	-0.01	-0.01	0.01	0.01	0.02
Retiring - Socially Bold	0.01	0.00	0.06	0.00	0.00	0.01	-0.02	0.00	-0.02	-0.01	0.00	0.00	0.00
Tough Minded - Tender Minded	0.01	0.04	0.03	0.02	0.01	0.00	0.00	0.06	0.00	0.13	0.11	0.12	0.04
Trusting – Suspicious	0.01	0.01	0.09	0.00	-0.02	0.00	0.03	0.01	-0.01	0.14	0.16	0.33	0.06
Practical – Abstract	0.01	0.00	0.05	0.00	0.00	0.01	0.03	0.00	0.01	-0.01	0.03	0.07	0.02
Forthright – Discreet	0.02	0.01	0.10	0.01	0.00	0.01	0.01	-0.02	0.01	0.02	0.03	0.03	0.02
Self-assured – Apprehensive	0.00	0.01	0.01	0.01	0.00	0.01	0.05	0.00	0.01	0.02	0.00	-0.03	0.01
Conventional – Radical	0.14	0.01	0.24	0.08	0.02	0.03	-0.01	0.04	0.03	0.60	0.54	0.54	0.19
Group - Orientated – Self-Sufficient	0.01	-0.01	0.02	0.00	0.00	0.00	0.00	0.03	-0.02	0.00	0.00	0.01	0.00
Undisciplined – Self-Disciplined	-0.01	0.00	0.24	0.00	0.00	0.01	-0.03	0.01	0.02	-0.06	-0.14	-0.01	0.00
Relaxed - Tense Driven	-0.29	0.02	0.21	-0.07	0.05	0.02	0.05	-0.18	0.06	-0.07	0.03	-0.09	-0.02
Average	-0.01	0.01	0.09	0.01	0.01	0.01	0.01	0.00	0.02	0.05	0.05	0.07	0.02

Item-level analyses

In analyses of variance of the item scores of the 15FQ+ adapted version, we found small levels of statistical significance both in the main effect of culture (uniform bias) and the interaction of culture and score level (non-uniform bias). Out of the 200 items, 60 (30%) turned out to be biased, which is a large proportion. However, the effect sizes ranged from 0.01 to 0.05, which points to a small effect size. It can be concluded that item bias is not a major disturbance in the 15FQ+ adapted version in these language groups. When comparing the item-bias analysis with the results of the original version (36% of the 200 items turned out to be biased), a slight decrease in the number of biased items is reported for the adapted version.

Influence of bias on size of cross-cultural differences. In order to inspect the impact of item bias on cross-cultural differences in the personality scales, the size of these differences was computed before and after the elimination of biased items. An item was taken to be biased if it had an eta square value of at least 0.02 for the uniform or non-uniform bias component, amounting to a total of 26 biased items (13%). One-way analyses of variance were carried out, with language group as independent variable and scale scores (sum scores on the items pertaining to the scale) as dependent variables. In a second step, the procedure was repeated, but now all biased items were excluded from the computation of scale scores. The extent of the cross-cultural differences was evaluated as the effect size (eta square) of the culture component. The effect size was 0.022 before the removal of biased items and 0.020 after bias removal. The same procedure of comparing effect sizes before and after the removal of biased items had been employed for the original (i.e., non-adapted) version of the 15FQ+. The effect size was 0.027 before the removal of biased items and 0.028 after bias removal). It can be concluded that the removal of biased items did not affect the size of the cross-cultural differences observed.

Discussion

The study addressed the cross-cultural suitability of an adapted version of the 15FQ+ in South Africa. Items of the original 15FQ+ were made easier to comprehend and culturally more appropriate for a diverse population. The instrument was administered in English to an applicant pool of 16,339 participants from all parts of South Africa who had applied for entry-level police jobs in the

South African Police Services. Bias was studied at construct and item level. The adapted version performed better than the original version. The structural equivalence of the adapted version was slightly higher. The four groups in which the first version showed fairly poor levels of structural equivalence (Ndebele, Asians, Coloureds, and Whites) showed higher levels of structural equivalence. The Relaxed – Tense factor remained problematic. Compared to the original version, internal consistency coefficients showed slight increases but for the Black group very low values were still reported.

The main question of the study is the effectiveness of the adaptation of the 15FQ+. The adaptations of the instrument essentially solved the cross-cultural equivalence problems, but low internal consistencies still make the instrument unsuitable for the South African context. Although it could be argued that the consistencies are not very low, given the small number of items in the scales (12 items per scale), the values are still so low that scale scores cannot be used as a basis of high-stake testing, such as hiring applicants. Furthermore, it is unlikely that further adaptations of the item contents would increase the internal consistencies in a major way.

The question has to be answered as to why in particular the Black group reports low alpha values for some of the scales. Despite the extensive adaptation of the 15FQ+, various items and words still seem to remain problematic for the Black groups. In a recent study by Wallice and Birt (2003), the understanding of the words (vocabulary) used in the 16PF (version SA92) was examined among English-speaking industrial psychology students. The students were instructed to provide synonyms for 135 words from the 16PF questionnaire. It was found that for the majority of the words participants in both language groups were unable to provide the correct synonyms. The groups indicated that they found it difficult to think of synonyms for words, although they apparently understood what the word meant. If the words were to be used in a sentence, it could have been easier to give their meanings. According to Wallice and Birt (2003) this finding suggests that the language in the 16 PF is still too difficult to be widely understood in a South African context.

There is some evidence that the size of the internal consistency of the 15FQ+ varies across socioeconomic strata of the South African society. Tredoux (2004) reports slightly higher alpha values for the 15FQ+ compared to the current study in a mixed-race sample of managers in a soft-drink manufacturing company, sales consultants in insurance industry, and professional and management candidates. It could well be that in Tredoux's study the linguistic and educational heterogeneity was smaller than in the current study. In a study among

senior police management, not yet published, the current authors also found slightly higher values than those found among the applicants. Still, the differences in internal consistencies across the studies were relatively small and none of the studies reported very high values. It can be concluded that the low internal consistencies of the current version of the instrument challenges its suitability for large-scale high-stakes assessment.

The question has to be addressed how we should proceed from here. Would additional refinements of the instrument further increase the cross-cultural equivalence of the measure? It could be speculated that further refinements of this instrument will only provide diminishing returns. The current adaptation procedure was extensive and was based on input from various informants with an expertise in the target cultures of the current study. Therefore, it seems fair to assume that if a major advancement of the cross-cultural suitability were possible with this instrument, the current authors would have identified the contours of the changes needed. Therefore, we tentatively conclude that our study points to the limits of the 15FQ+ and possible further adaptations in its current form for South Africa. It should be noted that the limits are more related to the low internal consistencies than to the poor equivalence of the scales. The current authors recommend that the factors that were problematic in the 15FQ+ adapted version (Accommodating – Dominant, Practical – Abstract, Self-assured – Apprehensive, Conventional – Radical, Relaxed – Tense Driven) need to be revisited on a construct level and items be redeveloped totally for the constructs. These adaptations need to take into account the cross-cultural sensitivities of the different groups in South Africa. As an alternative, future research could address new scoring keys (attempts have been done to develop a scoring key for a global factors) (Psytech, 2002). Also, higher-order constructs could be developed on the basis of the current 15 factors, which could show better internal consistencies because of their aggregated nature. Whatever approach is chosen, it should lead to an instrument that is quite different in its items, scales, or interpretations from the current version in order to meet the challenges of the Employment Equity Act.

Developing appropriate measures for the multicultural, multilingual South African society that comply with the Employment Equity Act seems to be problematic. Proper test usage assumes the availability of adequate instruments in a diverse society. Adequacy has to be demonstrated; are important. Special attention need to be given to standards put forward when testing individuals of diverse background, as was the case in this study. The current study illustrates various issues encountered when large-scale adaptations are implemented (the Standards for Educational and Psychological Testing, APA/AERA/NCME, 2000,

provide a good overview of issues to deal with). Although the goal of this study, namely, the development of a culturally appropriate version of the 15FQ+, was not realised, we would like to emphasize that low internal consistencies constituted the main reason for the failure. The bias study pointed to the adequacy of the adaptations. Therefore, it is concluded that the negative results were not due to a lack of a sound approach or methodology. Seen from this perspective, the current study may provide a template to emulate in the development or adaptation of other instruments.

References

AERA/APA/NCME. (2000). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of South Africa, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo and I. Ebersohn. *South African Journal of Psychology, 32,* 58–61.

Abrahams, F., & Mauer, K. F. (1999a). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology, 25,* 53–59.

Abrahams, F., & Mauer, K. F. (1999b). Qualitative and statistical impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South African context. *South African Journal of Psychology, 29,* 76–86.

Chan, W. Ho, R. M., Leung, K., Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378–402.

Claassen, N. C. W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47,* 297–307.

Gough, D., & England, V. (1991). *Medium of instruction in South Africa: Historical overview.* Unpublished report. Johannesburg: NEPI.

Government Gazette, Republic of South Africa, Vol. 400, no 19370. Cape Town, 19 October 1998.

Hambleton, R. J. (1994). Guidelines for adapting educational and psychological test: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

Hambleton, R. K. (2001). The next generation of the ITC Test Translation and Adaptation Guidelines. *European Journal of Psychological Assessment, 17,* 164–172.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.) (2004). *Adapting educational test and psychological test for cross-cultural assessment.* Mawhaw, NJ: Erlbaum.

Harkness, J. A. (2003) Questionnaire translation. In J. A. Harkness, F.J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York: Wiley.

Macdonald, C. A. (1990) *Crossing the threshold into standard three in Black education: The consolidated main project of the Threshold Project.* Pretoria: Human Sciences Research Council.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18,* 245–256.

Meiring, D. (2000, June). *Revisiting the cross-cultural comparability of the 16 Personality Factor Inventory (16PF) in the South African context.* Paper presented at the Industrial Psychology Conference (incorporating the Psychometrics Conference), Pretoria, South Africa.

Meiring, D., Van de Vijver, A. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality measures in South Africa. *South African Journal of Industrial Psychology, 31*(1), 1–8.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105–118.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398.

Nell, V. (1997). Science and politics meet at last: The insurance industry and neuropsychological norms. *South African Journal of Psychology, 27,* 34–49.

Owen, K. (1991). Test bias: The validity of the Junior Aptitude Tests (JAT) for various population groups in South Africa regarding the constructs measured. *South African Journal of Psychology, 21,* 112–118.

Prinsloo, C. H., & Ebersohn, I. (2002). Fair usage of the 16PF in personality assessment in South Africa: A response to Abrahams and Mauer with special reference to issues of research methodology. *South African Journal of Psychology, 32*, 48–57.

Psytech (2002). *The 15FQ+ Technical Manual.* Retrieved March 1, 2004 from http://www.15fq.com/Downloads/15fq+man.pdf.

Spence, B. A. (1982). *A psychological investigation into the characteristics of black guidance teachers.* Unpublished Master's dissertation, University of Pretoria, Pretoria, South Africa.

Shuttelworth-Jordaan, A. B. (1996). On not reinventing the wheel: A clinical perspective on culturally relevant test usage. *South African Journal of Psychology, 26,* 184–192.

Tact, H. (1999). *The cross-cultural validity and compatibility of the Sixteen Personality Factor Questionnaire.* Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

Taylor, I. A. (2000). *The construct comparability of the NEO PI-R Questionnaire for Black and White employees.* Unpublished doctoral thesis, University of the Orange Free State, Bloemfontein, South Africa.

Taylor, T. R., & Boeyens, J. C. A. (1991). *A comparison of black and white responses to the South African Personality Questionnaire.* Pretoria: Human Sciences Research Council.

Thissen, D. (1984). *MULTILOG: A user's guide.* Mooresville, IN: Scientific Software.

Tredoux, N. (2004). *15FQ+ SA Reliability.* Retrieved March 1, 2004 from http://www.psytech.co.za/15FQPlus_reliabilities.htm.

Tyler, G. (2002) *A review of the 15FQ+ Personality Questionnaire.* Pulloxhill, UK: Psychometrics Limited.

Van de Vijver, F. J. R. (2003) Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143–155). New York: Wiley.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

Van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (2nd ed., vol. 1) (pp. 257–300). Boston, MA: Allyn & Bacon.

Van de Vijver, F. J. R., & Leung, K. (1997b). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33,*141–156.

Wallice, T., & Birt, M. (2003). A comparison of native and non-native English-speaking groups' understanding of the vocabulary contained within the 16PF (SA92). *South African Journal of Psychology, 33,* 182–190

Chapter Four
Internal and External Bias of Cognitive and Personality Measures in South Africa*

Abstract

Internal and external bias was studied in two independent cohorts of entry-level applicants at the South African Police Service (N = 723 and 597) composed of three ethnic groups (Blacks, Coloreds, and Whites). Two cognitive tests and two personality questionnaires were used to predict training outcomes. We evaluated the test battery for compliance with the technical standards put forward in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures, developed in the U.S.A. and also adopted in South Africa. Neither cohort showed evidence of internal or external bias. Personality questionnaires were less powerful predictors than cognitive tests. Our study points to the cross-cultural generalizability of western findings regarding internal and external bias in selection procedures.

*Meiring, D., Van de Vijver, F. J. R., Rothmann, S., & Sackett, P. R. (2006). *Internal and external bias of cognitive and personality measures in South Africa.* (Manuscript submitted for publication).

Introduction

The current paper describes a study of internal and external bias in an assessment battery used by the South African Police Services to recruit entry-level police trainees. Internal (or measurement) bias focuses on the relationship between an observed score and a latent trait variable, whereas external (or predictive bias) focuses on the relationship between two observed variables – a predictor and a criterion (Saad & Sackett, 2002). Bias is an important issue in a multicultural society such as South Africa, as it may interfere with fair selection practices. The promulgation of the Employment Equity Act 55 of 1998, Section 8 in South Africa (Government Gazette, 1998) focuses specifically on this issue. According to this post-apartheid legislation,

> Psychological testing and other similar assessments are prohibited unless the test or assessment being used (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group

The Act imposes very strict criteria on psychological practice and only a few studies on internal bias and no studies on predictive bias in cognitive or personality measures have been carried out in South Africa. The Act creates a daunting task for psychology as a profession in South Africa for various reasons: there are virtually no relevant studies on internal and external bias, South Africa has a huge cultural and language diversity and there is a scarcity of valid psychological tests. For organizations in South Africa, the Act also requires the implementation of affirmative action measures to redress the disadvantages in employment experienced by three designated groups (i.e., Blacks, Coloreds, and Asians). Affirmative action is intended to ensure that suitably qualified employees from designated groups have equal employment opportunity and are equitably represented in all occupational categories and levels of the workforce. The Act specifies preferential treatment and numerical goals to ensure equitable representation.

Internal Bias in Cognitive and Personality Measures in South Africa

Internal bias (or structural equivalence) is often studied using exploratory or confirmatory factor analysis (Van de Vijver & Leung, 1997). Invariance of factor solutions across different cultural groups is usually seen as evidence for the absence of internal bias.

Few studies on the cross-cultural suitability of cognitive and personality tests had been carried out in South Africa before the 1980s (Meiring, Van de

Vijver, Rothmann, & Barrick, 2005). There has been an increasing interest in South Africa in the topic, thereby following international trends during the last few decades (Claassen, 1997). Studies in South Africa have reported race, education, language, and understanding of English as sources of incomparability of constructs and items of cognitive and personality tests (Abrahams, 1996, 2002; Abrahams & Mauer, 1999a, 1999b; Meiring, 2000; Meiring, Van De Vijver, & Rothmann, 2006); Meiring, Van De Vijver, Rothmann, & Barrick, 2005; Wallice & Birt, 2003).

Our research group has examined the internal bias of the selection battery employed by the South African Police Services (SAPS), which consisted of an intelligence and a personality measure. The cognitive instruments showed structural equivalence and almost no item bias (Meiring et al., 2005). Several scales of the personality questionnaire showed low internal consistencies and a poor structural equivalence across ethnic and linguistic groups. The item bias in the personality scales was low. Method bias (a generic term for all instrument-related factors that invalidate cross-cultural comparisons, such as social desirability, Van de Vijver & Leung, 1997) did not influence the size of the cross-cultural differences in the personality scales. In addition, several personality scales revealed low internal consistencies, notably in the Black groups. In a second study (Meiring et al., 2006), the personality instrument was adapted in order to improve the structural equivalence and the internal consistencies of the scales (mainly by simplifying the language of various items). The structural equivalence in the adapted 15 FQ+ improved and was adequate for most scales in most groups but low internal consistencies continued to limit the usefulness of the inventory.

External Bias in Cognitive and Personality Measures

Useful criteria for determining predictive bias are proposed in both the Standards for Educational and Psychological Testing (American Educational Research Association/American Psychological Association/National Council on Measurement in Education, (AERA/APA/NCME), 1999), the fourth edition of the Principles for the Validation and Use of Personnel Selection Procedures (Society of Industrial and Organizational Psychology, 2003) and the Guidelines for the Validation and Use of Assessment Procedures for the Workplace (Society for Industrial and Organizational Psychology in South Africa, 2005). The first criterion is that an analysis of predictive bias requires an unbiased criterion. Confidence in the criterion measure is a prerequisite for the analysis of predictive bias. The second criterion is the issue of statistical power to detect slope and intercept

differences. Small total or subgroup sample sizes, unequal subgroup sample sizes, range restriction, and predictor unreliability are factors contributing to low power. The third criterion is the assumption of homogeneity of error variances of the criterion scores. Equal slopes and intercepts are the main criterion for the absence of predictive bias.

To our knowledge, no South African study on predictive bias in cognitive or personality measures has been reported in the literature. However, there is much literature on differential prediction mostly on race and gender in other countries (e.g., AERA/APA/NCME,1999; Hough, Oswald, & Ployhart, 2001; Neisser et al., 1996; Sackett & Wilk, 1994; Wightman, 1997). Cognitive ability measures are among the most valid predictors of job performance across a wide variety of jobs, as confirmed by various meta-analysis studies (Alonso, Viswesvaran, & Sanchez, 2001, Hunter, 1980, Hunter & Hunter 1984, Pearlman, Schmidt, & Hunter, 1980; Schmidt & Hunter, 1981, 1998, 2004). The American literature shows fairly consistent differences of 0.7 *SD* to 1 *SD* between ethnic groups, with Asian Americans obtaining the highest scores followed by European Americans and Hispanics and African Americans (e.g., Hunter, 1986). Studies of predictive validity address the question to what extent the difference in cognitive test performance that is frequently found in the cognitive domain is also reflected in criterion scores. Most studies have been carried out in the U.S.A. Slope differences are rarely found (Bartlett, Bokko, Mosier, & Hanna, 1978; Hunter, Schmidt, & Rauschenberger, 1984; Schmidt, Pearlman, & Hunter 1980). When differential prediction is found, it usually concerns differences in intercepts, in which case the regression line of the majority group is almost always above the regression line of the minority group. In other words, in most cases it turns out that cognitive measures are unbiased predictors and when predictive bias is found, it is to the advantage of the minority group.

Personality as a predictor of job performance in the workplace has recently received a lot of attention in the literature (e.g., Hogan 2005; Hough 2001, Ones, Viswesvaran, & Dichert 2005). The use of the Five-Factor model has given much momentum to researchers' understanding of personality-based predictor relationships with work performance (Barrick, Mount, & Judge, 2001; Hogan & Holland 2003; Hurtz & Donovan, 2000; Salgado, 1997). Ones, Viswesvaran, and Dichert (2005), after analyzing various meta-analytic investigations, reported substantial validity coefficients of .28 for the Big Five personality dimensions, Operational validities for overall job performance and its various facets were in the .30s, but as high as .47 for teamwork, .44 for training performance, and .43 for citizenship performance. Validities for Big Five

personality dimensions as a set are substantial according to Ones et al. Hogan (2005) indicates that in adequately conducted studies the correlations between standard dimensions of normal personality and job performance criteria that are relevant to these dimensions are reliably above .30 while multiple correlations approach .50. He argues that personality predicts occupational performance almost as well as measures of cognitive ability. Unlike cognitive ability measures, personality measures do not discriminate in terms of race and do not show huge group differences. Work on differential prediction of personality questionnaires is now beginning to appear (De Meijer, Born, Terlouw, & Van der Molen, 2006, Te Nijenhuis & Van der Flier, 2000; Saad & Sackett, 2002). For instance, de Meijer et al. 2006 study examined the criterion-related validity of both cognitive and non-cognitive selection measures on training success of Dutch police trainees. They reported that ethnic score differences on the cognitive ability test were larger than ethnic score differences on non-cognitive selection measures and in training results. Roughly, training success was better predicted by cognitive ability for the ethnic majority group, while training success was better predicted by non-cognitive instruments for the ethnic minority trainees. Nijenhuis and Van der Flier (2000) report that differential prediction only plays a role for the less cognitive and less objective criteria. Saad and Sackett (2002) reported findings parallel to those in the ability domain in examining predictive bias by gender using personality measures (i.e., little evidence of slope differences and intercept differences in the form of over prediction of female performance). This literature does not show a consistent picture; there is a need for more comprehensive analysis of predictive bias on personality measures.

Research Aims

The present study has three research aims; firstly, we examined predictive bias in two independent cohorts of entry-level applicants at the South African Police Service (in 2000 and 2001). Two cognitive tests (developed by the South African Police Services) and two personality instruments (the Fifteen Factor Questionnaire, 15FQ+, (Psytech 2002a), and the Occupational Personality Profile, OPP, (Psytech, 2002b) were used to predict training outcomes. We addressed the question to what extent the battery complies with the technical standards put forward in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures (Society of Industrial and Organizational Psychology, 2003), Secondly, we were interested in the cross-cultural generalizability of findings of western predictive bias studies to South Africa. Predictive bias in cognitive and personality measures as found in South Africa is

compared with western-based studies reported in the literature. Thirdly, we were interested in the suitability of the SIOP guidelines for South Africa, in which larger cross-cultural differences exist than in the countries (such as the U.S.A.) in which the guidelines have been employed before.

STUDY 1

Method

Participants

The group consisted of 723 selected applicants from the original group of 13,681 who applied for entry-level police positions. The sample consisted of 152 women and 571 men. The age varied from 18 to 36 years, with a mean of 24.74 years and (*SD* = 2.86). The sample was comprised of 468 Blacks, 123 Coloreds, and 132 Whites1. A successfully completed "matric" exam (comparable to Grade 12) is an entry-level requirement; 59.3% of the applicants had this qualification, 10.8% had a one- to two-year diploma, 5.7% had a three-year degree, and 16.4% had a post-graduate qualification.

Measurement and Procedure

Predictors

All the measuring instruments were administered in English.

Cognitive measures. The two cognitive tests (reading and comprehension test, spelling test) were developed specifically for the South African Police Service (Meiring, Van de Vijver, Rothmann, & Barrick, 2005). The reading and comprehension test consisted of four paragraphs that were selected from the basic training modules (Module 1: the Bill of Rights on Police Power, Community Policing; Module 2: Non-Verbal Communication; Module 5: Mental Disorder). The spelling test was also developed for the South African Police Service. Training instructors at the training college were asked to generate a pool of police-relevant words (such as *rape* and *homicide*). A pool of 40 words was generated.

Personality inventories. The 15FQ+ is a 200-item personality test with a trichotomous response scale, developed by Psytech International (Psytech, 2002a). The authors' aim was to produce a relatively short, yet robust measure of Cattell's primary personality factors. The 15FQ+ measures the following dimensions: Cool Reserved – Outgoing; Intellectance; Affected by Feelings -

---

1 We adopted the terms Blacks, Coloreds, and Whites here, because these terms are commonly used in South Africa, both in public discourse and in scientific writings. The sample also had a subgroup of Asians, which was not considered here because of its limited size.

Emotionally Stable; Accommodating – Dominant; Sober serious – Enthusiastic; Expedient – Conscientious; Retiring – Socially Bold; Tough Minded - Tender Minded; Trusting – Suspicious; Practical – Abstract; Forthright – Discreet; Self-assured – Apprehensive; Conventional – Radical; Group - Orientated - Self-Sufficient; Undisciplined - Self-Disciplined; Relaxed – Tense Driven. The internal consistencies of the 15FQ+ are favorable with mean values of above .70 for most of the scales according to the manual (Psytech, 2002a). Tylor (2002) reports evidence supporting the construct validity of the 15FQ+ as measured by correlations with other personality measures (16PF, 16PF5, and five-factor model).

The Occupational Personality Profile (OPP), also developed by Psytech International (2002b), is a personality measure for use in industrial and organizational settings. The inventory consists of 98 items and uses a five point-Likert response format ranging from *strongly agree* (1) to *strongly disagree* (5). The OPP measures nine different personality dimensions (accommodating - assertive, detail-conscious – flexible, cynical –trusting, emotional – phlegmatic, reserved – gregarious, genuine - persuasive, composed – contesting, optimistic – pessimistic, abstract - pragmatic) in addition to the distortion scale. Reliability coefficients above .60 have been reported for all constructs of the OPP; test re-test reliability. The manual of the OPP reports favorable results for both the construct and predictive validity of the OPP (Psytech, 2002b).

Criterion measure. The basic training program of the South African Police Service was used as the criterion measure. This consists of a modular program that needs to be successfully completed over a period of six months, followed by weapon training. The criterion measures were comprised of the assessment of the training outcomes of the following ten modules: professional police conduct; verbal and written communication in the community service centre with reference to crime reporting; preventive policing and the rendering of a professional service to the community; reactive policing and the approach to the scene of a crime; searching, seizure and forfeiture; different crimes; vehicles and related matters; detention; law of evidence; and self-management. The internal consistencies for the different ethnic groups on the criterion measure, which assessed knowledge of all domains covered in the training were high (Blacks .90, Coloreds .90, and Whites .86).

Procedure. Predictor and criteria data were collected from three cohorts. A two-stage selection strategy was followed, in which applicants had to obtain a certain cut-off score on the cognitive criteria after which they had to comply with a personality profile on the personality measures. The applicants had to

successfully complete all the prescribed modules during the six months of basic training. The groups were randomly divided into platoons consisting of 34 to 38 members each. The group composition of the platoons was random with regard to ethnic identity. Each platoon had its own instructor who was responsible for the complete training. Training and assessment in the 10 modules followed standardized procedures. The knowledge and understanding of the recruits were assessed by means of a standardized test at various points in the training program. The percentage of the items correctly answered was calculated for each of the modules and at the end of the six months training period and an average overall mean score was determined (mean criterion score). Recruits were also assessed by the platoon instructor. The mean criterion scores were standardized per instructor in order to avoid interrater differences.

Statistical analysis. The analysis involved six steps. We *first* considered the internal bias (structural equivalence) of both the predictors and the criterion. The internal bias of the predictors in the unselected and much larger group applicants has been studied before (Meiring et al., 2005, 2006). Factors obtained in the three cultural groups were compared to the factors obtained in the pooled group, in which all data were combined (Muthén, 1991, 1994). Factors in the cultural groups were rotated to the factors of the pooled solution. The agreement was evaluated by means of a factor congruence coefficient, Tucker's phi (Chan, Ho, Leung, Cha, & Yung, 1999; Van de Vijver & Leung, 1997a, b). Values above .90 are taken to point to essential agreement and values above .95 to very high agreement. The *second* step involved the estimation of the internal consistencies of cognitive, personality and criteria scales.

The *third* step dealt with the internal consistencies of the personality measures. In two previous studies involving the unselected sample from which the current sample is recruited, low internal consistencies of the predictors were found in all groups (Meiring et al., 2005, 2006). It is likely that the sample of selected applicants also shows unacceptably low internal consistencies.

An obvious way to deal with low internal consistencies would be to aggregate items or scales of the two personality inventories. A problem in aggregation might be the presence of method factors, as the two inventories did not use the same response formats (the 15FQ+ employs a three-point response scale and the OPP a five-point scale). Interbattery factor analysis has been proposed as a tool that can overcome method factors (Tucker, 1958; see also Browne, 1979). We employed maximum-likelihood interbattery factor analysis (Browne, 1979) to identify those factors that are common to the 15FQ+ and OPP variables (labeled superfactors in the remainder). The technique identifies

dimensions that are common to the underlying factor structure of the two instruments. This method is an appropriate alternative to exploratory factor analysis due to its ability to separate within-method and cross-method variability (Cudeck, 1982).

The *fourth* step addressed the size of group differences. Prior to the analysis, a factor analysis of the two cognitive tests was carried out in order to combine their scores in a single score. The first factor explained 72% of the variance. The size of the group differences was tested in a one-way multivariate analysis of variance with cultural group as independent variable and scores on the superfactors (personality), cognitive test, and criterion as dependent variables.

The *fifth* step tested for differential validity (predictive bias) using multigroup structural equation modelling. The AMOS program (Arbuckle, 2003) was used to test cross-cultural differences of the slopes and intercepts in each group by relating the independent variables (scores of the cognitive factor and the personality superfactors) to the dependent variable (scores on the training tests). The model we used only contained observed variables (apart from the error variance of the dependent variable). Independent variables were allowed to correlate. A set of hierarchically nested regression models was tested. The standard sequence of parameter constraints was used: null model (no constraints on the identity of parameters across cultures), equal regression coefficients ("structural weights", also referred to as "equal slopes"), equal intercepts ("structural intercepts", also referred to as "equal intercepts"), equal means of the predictors ("structural means"), equal covariances of the predictors ("structural covariances"), and equal error terms of the dependent variable ("structural residuals") (Byrne, 2001). The fit of each model was examined by examining the significance of the increments in chi-square values.

In the *sixth* step we tested the influence of unreliability and range restriction. The correction was based on the availability of all predictor scores in the full, unselected sample. We computed the correlation between each of the predictors and the criterion score in the sample of the selected applicant group and corrected this correlation for range restriction using the standard deviation in the selected group and the unselected sample.

Results

Internal Bias

The first step involved the evaluation of structural equivalence. Based on scree tests, the reading and comprehension test of the cognitive predictor showed

a unifactorial solution in the pooled data across the three different groups. Table 4.1 shows that for the spelling test the Colored group had a value of Tucker's phi lower than .90. The mean agreement of the unifactorial scales for both the 15FQ+ and OPP showed Tucker's phi values higher than .90; yet, various scales, notably of the 15FQ+, showed much lower values. Finally, factor analyses of the criterion scores yielded a one-factorial solution, which showed high agreement values. It can be concluded that there is no evidence for internal bias of the cognitive and criterion measures in these samples, while the personality scales showed more ambiguous results.

Internal consistency. The second step examined the internal consistencies of all instruments. Table 4.1 reports the internal consistency coefficients for the cognitive, 15FQ+, OPP, superfactors, and criterion measures. Results from Table 4.1 indicate moderate levels of internal consistencies for the cognitive measures, whereas the 15FQ+ and the OPP revealed low mean values. The criterion measure showed a high internal consistency.

Interbattery factor analysis. An interbattery factor analysis was carried out in the third step. Three-, four-, and five-factor solutions were obtained. The Varimax rotated five-factor solution provided the most parsimonious and theoretically meaningful solution. Moreover, the Tucker-Lewis index for the three-, four-, and five-factor solutions were .69, .85, and .91, respectively. The loadings are presented in Table 4.2. The first factor, labeled Extraversion, showed the highest loadings for Gregarious (.81), Enthusiastic (.68), and Self-Sufficient (.72). The second factor, labeled Dominance, was defined mainly by Assertive (.64), Persuasive (.60), and Dominant (.60). The third factor was labeled Stability and the following scales loaded on it: Phlegmatic (.71), Emotionally Stable (.59), and Self-Assured (-.54). The fourth factor was labeled Openness and showed high loadings for Pragmatic (.75) and Tender Minded (.70). The fifth and last factor, labeled Agreeableness, showed the highest loadings for Cynical - Trusting of the OPP (.70) and Trusting – Suspicious of the 15FQ+ (-.67).

In order to compute the internal consistency of the superfactors, (only) the items on the italicized scales in Table 4.2 were used. As can be seen in Table 4.1, the internal consistencies were high. The interbattery factor analysis suggested useful groupings of the subscales of the two personality inventories with high internal consistencies. Furthermore, the superfactors showed high levels of structural equivalence; the agreement of the superfactors in the pooled solution with factors in the three race groups were well above the threshold value of .90 (see Table 4.1). It can be concluded that the interbattery factor analysis

addressed the problems with the internal consistencies and the problematic structural equivalence of the subscales of the two personality questionnaires.

External Bias

Size of group differences. The effect size of cross-cultural differences for the predictor and criterion variables are reported in Table 4.3. All personality factors showed significant differences ($p < .01$). Large effect sizes were found for extraversion ($\eta^2 = .15$) and openness (.20), a medium size effect was found for agreeableness (.08) and small differences were found for dominance (.03) and stability (.01). The scores of Table 4.3 were standardized in such a way that across the whole sample, the mean was zero ($SD = 1$), so that scores in the Table can be interpreted as z score differences. Compared to the other groups, Blacks scored lowest on extraversion, openness, and agreeableness. The means of the Whites and Coloreds were close to each other on all scales. There were no significant intergroup differences in means on cognitive skills and criterion scores; this result could be expected in a two-stage procedure in which relatively high scores on the cognitive instruments are a prerequisite for entering the second stage.

Table 4.1 *Factor Agreement Indexes and Internal Consistencies per Cultural Group (Study 1)*

	Factor agreement			Internal consistencies		
	Blacks	Coloreds	Whites	Blacks	Coloreds	Whites
Cognitive						
Spelling	.99	.85	.93	.78	.68	.78
Reading	.98	.90	.94	.69	.68	.79
Comprehension						
Personality						
Fifteen Factor Questionnaire 15FQ+						
Mean	.92	.90	.93	.33	.27	.28
Min	.50	.75	.65	.08	.05	.03
Max	.99	.99	.99	.50	.53	.58
Occupational Personality Profile						
Mean	.96	.94	.95	.46	.44	.45
Min	.77	.90	.91	.30	.22	.05
Max	.99	.98	.99	.62	.69	.74
Superfactors						
Extraversion	.98	.98	.97	.73	.84	.84
Dominance	.97	.93	.97	.61	.75	.81
Stability	.97	.97	.93	.73	.81	.81
Agreeableness	.97	.95	.89	.60	.67	.72
Openness	.89	.97	.89	.54	.59	.68
Criterion	.99	.99	.99	.87	.86	.90

Table 4.2 *Results of the Interbattery Factor Analysis (Study 1)*

Personality Scales	Extraversion	Dominance	Stability	Openness	Agreeable-ness
Fifteen Factor Questionnaire 15FQ+					
Cool Reserved – Outgoing	.31	-.02	-.18	-.28	-.01
Intellectance	-.13	.25	.44	.16	-.05
Affected by Feelings – Emotionally Stable	-.10	.00	*.59*	.05	.03
Accommodating – Dominant	.02	*.60*	.26	.05	-.12
Sober Serious – Enthusiastic	*.68*	.20	.06	-.03	.09
Expedient Conscientious	.14	.18	.21	.17	-.24
Retiring – Socially Bold	-.28	.51	.28	.24	.03
Tough Minded – Tender Minded	-.03	-.02	.00	*.70*	.03
Trusting – Suspicious	.21	.09	-.22	-.01	*-.67*
Practical – Abstract	.03	.27	-.17	.22	-.04
Forthright – Discreet	.06	-.34	.10	.30	-.08
Self-assured – Apprehensive	.05	-.14	*-.54*	.08	-.18
Conventional – Radical	-.08	.26	-.08	-.03	.08
Group Orientated – Self-Sufficient	*.72*	-.02	-.11	-.02	-.12
Undisciplined – Self-Disciplined	-.15	-.08	.17	-.08	-.14
Relaxed – Tense Driven	.07	.22	-.44	-.15	.05
Occupational Personality Profile					
Accommodating-Assertive	.08	*.64*	.15	-.08	-.10
Detail-Conscious–Flexible	.00	-.02	.08	.00	*-.33*
Cynical – Trusting	-.09	-.04	.27	.02	*.70*
Emotional – Phlegmatic	-.13	.09	*.71*	-.04	.08
Reserved – Gregarious	*.81*	-.04	-.16	.00	-.16
Genuine – Persuasive	-.17	*.60*	.04	.24	-.11
Composed – Contesting	.15	.28	-.15	.16	-.36
Optimistic – Pessimistic	.06	-.03	-.43	-.03	-.32
Abstract – Pragmatic	.00	.18	.07	*.75*	-.02

Note. Strongest loadings in italics.

Table 4.3 *Means of the Standardized Scores on the Scales of the Superfactors, Cognitive Skills and Criteria per Ethnic Group (Study 1)*

Scale	Blacks	Coloreds	Whites	η^2
Extraversion	-0.24	0.08	0.78	.15**
Dominance	0.12	-0.20	-0.22	.03**
Agreeableness	-0.18	0.05	0.59	.08**
Stability	-0.08	0.18	-0.12	.01**
Openness	-0.32	0.51	0.69	.20**
Cognitive Skill	0.01	-0.07	0.03	.00
Criterion	0.01	0.12	-0.16	.01

Note. The last column gives the partial eta square value which indicates the proportion of variance accounted for by cultural group differences in the analysis of variance.

Multigroup structural equation modeling. The results for the regression analysis are reported in Table 4.4. The model fit did not significantly decrease by introducing an equality constraint on the regression coefficients. The introduction of equal intercepts showed a marginally nonsignificant effect ($\Delta\chi^2(2) = 5.22$, $p = .07$). Introducing cross-cultural equality constraints on both the means of the predictors and their covariances yielded highly significant increments of the chi square statistic. Finally, holding the error term of the dependent variable identical across cultures did not lead to a significant increase of the fit. It can be concluded that there was unambiguous support for the cross-cultural identity of the slopes and the error variances of the dependent variable and equally unambiguous support for the heterogeneity of predictor means and covariances; identity of intercepts was not clearly supported (nor could it be easily rejected).

Table 4.4 *Fit Statistics of the Multigroup Regession Analysis: Increments in Chi Square Statistics and Their Level of Significance (Study 1)*

Model	$\Delta\chi^2$	Δdf	*P*
Structural weights	4.43	10	.93
Structural intercepts	5.22	2	.07
Structural means	203.06	12	.01
Structural covariances	115.81	30	.01
Structural residuals	2.37	2	.31

The estimated parameters of the model with equal slopes and intercept are presented in Table 4.5. Significant predictors (with a positive regression weight) were extraversion, stability, and cognitive skill. The proportion of variance explained by the predictors was significant, though small; the multiple correlations were .15 (R^2 = .022) for the Black group, .18 (R^2 = .030) for the Colored group, and .15. (R^2 = .024) for the White group.

Table 4.5 *Parameters of the Regression model with Equal Slopes and Intercepts in Each Cultural Group (Study 1)*

Predictor	*B*	β	*P*
Extraversion	0.64	0.07	.01
Dominance	0.56	0.03	.48
Agreeableness	0.15	0.01	.83
Stability	3.13	0.14	.01
Openness	-0.59	-0.03	.40
Cognitive Skill	0.64	0.07	.01
Slope	50.05		

Correction for unreliability and range restriction. The effects of unreliability and range restriction are reported in Table 4.6. The variance ratio (of score variance in the selected sample to the total sample) was larger than one for all variables, except for Stability. The large variance ratios are not in line with the literature. Particularly the large ratio of Cognitive Skill was unexpected, as high sores on this variable are needed for selection. The selection strategy specifies that in order to be accepted, an applicant should have high score on Cognitive Skill, but at the same time the applicant should have a certain personality profile. So, sufficient scores on Cognitive Skill, which were obtained by a large subsample, do not yet guarantee a positive selection outcome. The mean scores in the selected group were higher in the selected group than in the total sample (63.71 and 58.65, respectively). The scores on Extraversion were somewhat lower in the selected sample. The most important predictors were Cognitive Skill and Stability. Corrections for unreliability and range restriction had only a minor influence on the correlations between the predictors and the outcome. Cognitive Skill and Stability remained the most salient predictors. The selection ratios were very dissimilar for the three groups; the ratio was .02 for the Blacks, .10 for the Coloreds, and .12 for the Whites.

Table 4.6 *Effect of Unreliability and Range Restriction on Correlation between Predictors and Criterion (Study 1)*

Predictor	*M* and *SD* in total sample group *(N = 13528)*	*M* and *SD* in Selected Group *(N = 423)*	Variance ratio	Correlation in selected group	Estimated correlation in total sample
Extraversion	0 (0.22)	-0.09 (0.22)	1.01	-.02	-.03
Dominance	0 (0.24)	0.03 (0.25)	1.10	.02	.02
Agreeableness	0 (0.28)	0.00 (0.28)	1.10	-.04	-.04
Stability	0 (0.21)	-0.02 (0.18)	0.78	.13**	.22
Openness	0 (0.28)	-0.04 (0.29)	1.06	-.02	-.03
Cognitive Skill	58.65 (13.22)	63.71 (14.46)	1.20	.13**	.14

*$p < .05$. **$p < .01$.

Discussion

The study attempted to examine to what extent the battery that is employed by the South African Police Service to recruit new officers complies with the technical standards put forward in the fourth edition of the Principles for the Validation and Use of Personnel Selection vis-à-vis internal and external bias (Society of Industrial and Organizational Psychology, 2003; Society for Industrial and Organizational Psychology in South Africa, 2005). The first is the use of an unbiased criterion. The criterion score was a composite of various training modules. The reliability coefficient of the scale in each ethnic group and the strong support for the structural equivalence of the criterion score did not reveal evidence of bias in the criterion. The second standard deals with the issue of statistical power to detect slope and intercept differences. Small total or subgroup sample sizes, highly unequal subgroup sample sizes, range restriction, and predictor unreliability are factors contributing to low power. The group of selected applicants consisted of 753 trainees representing three of the four ethnic groups

(the Asian group was not included due to its small sample). Predictor reliability is a third important consideration. The cognitive tests showed fair reliabilities in the different race groups. However, the scales of the personality measures showed unacceptably internal consistencies. An interbattery factor analysis resolved the problem by establishing a new structure with five scales, which showed good internal consistencies. The assumption of homogeneity of error variances was also examined as a third requirement. The structural equation model did not show a significant decrease of fit when the error variances of the dependent variables were kept constant across ethnic groups. A good fit was found for a regression model with an equal slope and intercept in each group. It can be concluded from the above that we complied with the technical standards and that we did not find evidence for the presence of predictive bias. The statistical core in determining differential prediction is the test of identity of slopes and intercept across the different cultural groups. The results revealed unambiguous support for the cross-cultural identity of both the slopes and intercepts, although the values of the squared multiple correlations were small, with only a few predictors being significant (e.g., extraversion, stability, and cognitive skill). Contrary to our expectation, we did not find a sizeable influence of range restriction on the regression results. Range restriction did not have a major impact on the variance of any variable, presumably because the selection is based on a combination of cognitive abilities and personality traits. Finally, the selection ratios were much more favourable for the Whites than for the other groups, presumably due to a variety of factors such as educational differences; we return to this issue in the General Discussion.

The results of the first study concur with findings in the literature in which cognitive ability predicted training performance (Hough & Oswald, 2000; Salgado et al., 2003; Schmidt & Hunter 1989, 2004). Stability contributed also to the prediction of training outcomes. Furthermore, we found no significant differences on the cognitive tests and highly significant differences on the personality measures. A possible explanation can be that the strong selection may have created equal groups of Blacks, Coloreds, and Whites.

STUDY 2

Although the findings about the predictive validity of the test battery were encouraging, we assumed that further improvements were possible, such as more explained variance and a variance reduction after selection that would be more in line with the literature. Therefore, we carried out an extended replication of the first study. The first change involves the use of adapted personality instruments. We attempted to reduce the internal bias by simplifying the language of presumably difficult items; more information about the adaptations can be found in Meiring et al. (2006). We were interested in the question of whether these adaptations would increase the predictive power of the personality instruments and would affect differential bias. The second change involves the analyses; the effects of range restriction and unreliability of the predictor and criterion scores are studied.

Method

Participants

The 2001 sample consisted of 597 entry-level students from the original group of 16,005 applicants who applied for entry-level positions in the South African Police Service. The sample consisted of 144 women and 453 men. The age varied from 18 to 30 years with a mean of 24.51 years (SD = 3.10). The ethnic composition was mixed; there were 365 Blacks, 90 Colored and 142 Whites. 60.3% of the applicants had grade 12 qualification, 14.9% had grade 12 with university exemption, 9.5% had a one- or two-year diploma (of tertiary education), 11.9% had a three-year degree or diploma, and 1.7% had a post-graduate qualification.

Measurement and Procedure

Predictors

Cognitive Measures. The cognitive tests of the first study were also applied here.

Personality Measures. Adapted versions of the 15FQ+ and the OPP employed in the first study were administered. The adaptation of both the 15FQ+ and the OPP followed a committee approach outlined by Harkness (2003) and Van de Vijver and Leung (1997). A group of experts revised many items in order to make them culturally and linguistically more appropriate for the South African context. A total of 44% (85 items) of the original 15FQ+ items were changed; in 38 items, one or two words were changed and in 47 other items the item stem

was reworded. The changes mainly involved three constructs (Practical - Abstract, Conventional - Radical, Relaxed - Tense Driven); major changes (rephrasing of sentences) were implemented in 21 items of these scales (out of a total of 36 items) while minor changes (changing one or two words) were introduced in 11 items. For the OPP, a total of 43% (42 items) of the original items were changed (about half of these changes were major). Most changes were made to five constructs: Accommodating – Assertive, Detail-Conscious – Flexible, Cynical – Trusting, Reserved – Gregarious, and Abstract - Pragmatic; 17 items of these scales (out of a total of 50 items) were rephrased and new words were used in 14 items.

Criterion. The criterion measures used in Study 1 were also employed in study 2, with exception of Module 7 (vehicles and related matters) and Module 10 (self-management), which were not used in the current study.

Procedure. The procedure for collecting criterion data was the same as in the previous study. The same two-stage selection strategy was employed again. Data were collected from three intake groups at the police college (June 2001, September 2001, and February 2002).

Statistical Analysis. The analysis followed the same steps as in the first study.

Results

Internal Bias

As can be seen in Table 4.7, we found strong evidence for the lack of internal bias in the cognitive and criterion measures, with a minor exception for reading and comprehension for the White group which showed a boundary value of .89. The scales of both the 15FQ+ and the OPP showed higher factor agreement indices than in the previous studies; yet, some subscales showed values well below the threshold value.

Table 4.7 *Factor Agreement Indexes and Internal Consistencies per Cultural Group (Study 2)*

	Factor agreement			Internal consistencies		
	Blacks	Coloreds	Whites	Blacks	Coloreds	Whites
Cognitive						
Spelling	.99	.96	.97	.70	.75	.76
Reading	.98	.97	.89	.60	.67	.80
Comprehension						
Personality						
Fifteen Factor Questionnaire 15FQ+						
Mean	.95	.93	.93	.48	.61	.68
Min	.81	.83	.59	.07	.45	.50
Max	.99	.98	.99	.70	.72	.82
Occupational Personality Profile						
Mean	.96	.93	.96	.53	.53	.55
Min	.80	.78	.91	.29	.36	.26
Max	.99	.99	.91	.66	.66	.80
Superfactors						
Extraversion	.98	.83	.97	.77	.72	.84
Dominance	.97	.96	.97	.62	.80	.83
Stability	.95	.93	.95	.75	.83	.90
Agreeableness	.98	.95	.98	.75	.80	.84
Openness	.98	.98	.96	.77	.83	.80
Criterion	.99	.95	.98	.62	.78	.80

Internal consistency. The second step examined the internal consistencies of all instruments. Table 4.7 reports the internal consistencies for the cognitive, 15FQ+, OPP, superfactors, and criterion measures. As can be seen in Table 4.7, moderate to high internal consistencies of the cognitive measures were found, which replicated the findings of the previous study. Both the 15FQ+ and the OPP showed higher mean values than in the first study; across cultural groups, the mean value for the 15FQ+ was .59 (an increase of .30) and the mean for the OPP was .53 (an increase of .08). The adaptation of the personality instruments had a positive effect on reliability coefficients, although unacceptably low values were again found for some scales. Finally, the criterion measure showed a good internal consistency (of .73), although this value was lower than the .88 found in the first study.

Interbattery factor analysis. Again, a five-factor solution yielded an interpretable solution with an acceptable value of the Tucker-Lewis index (value for the three-, four-, and five-factor solutions were .72, .86, and .92, respectively). Factor loadings are presented in Table 4.8. The first factor was labeled Extraversion; the scales with the highest loadings were Gregarious (.78), Enthusiastic (.68), and Self-Sufficient (.65). The second factor, labeled Dominance, was marked by Assertive (.71) and Dominant (.65). The third factor was labeled Stability and showed the highest loadings for Phlegmatic (.68) and Emotionally Stable (.59). The fourth factor, labeled Openness, was mainly defined by Pragmatic (.77) and Tender Minded (.69). Finally, the fifth factor represented Agreeableness; the scales with the strongest loadings were Trusting (.74) and Suspicious (-.73). The superfactors showed higher levels of internal consistencies than in the previous study. The average increases were .09 for the Black group, .06 for the Colored group, and .07 for the White group. Strong support was found for the structural equivalence of the superfactors, with a single exception: the agreement index of Extraversion was .83 in the Colored group. It was decided to retain the factor in this group because the patterning of the loadings did not point to a different psychological meaning for the factor.

Table 4.8 *Results of the Interbattery Factor Analysis (Study 2)*

	Factors				
Personality Scale	Extraversion	Dominance	Stability	Openness	Agreeableness
Fifteen Factor Questionnaire					
Cool Reserved – Outgoing	.35	.00	.09	.26	.01
Intellectance	.22	.21	.34	.36	-.10
Affected by Feelings – Emotionally Stable	.08	.04	*.59*	.05	.25
Accommodating – Dominant	.14	*.65*	.22	.16	-.05
Sober serious – Enthusiastic	*.68*	.14	-.09	.16	.12
Expedient – Conscientious	-.03	.06	.31	.21	-.04
Retiring - Socially Bold	.49	.22	.42	.24	.00
Tough Minded - Tender Minded	.02	-.11	-.04	*.69*	.10
Trusting – Suspicious	-.19	.10	-.07	-.09	*-.73*
Practical – Abstract	.15	.09	-.03	.30	-.03
Forthright – Discreet	.08	-.36	-.21	.11	.04
Self-assured – Apprehensive	-.05	-.09	-.44	.09	-.10
Conventional – Radical	.10	.16	-.21	-.07	-.10
Group – Orientated - Self-Sufficient	*.65*	-.08	.03	-.01	.21
Undisciplined – Self-Disciplined	-.06	-.02	.20	-.01	-.10
Relaxed - Tense Driven	-.06	.03	-.32	.01	-.14
Occupational Personality Profile					
Accommodating – Assertive	.09	*.71*	.15	.08	-.06
Detail-Conscious – Flexible	.02	.03	-.19	-.08	.14
Cynical – Trusting	.17	-.10	.16	.06	*.74*
Emotional – Phlegmatic	.22	.09	*.68*	.02	.23
Reserved – Gregarious	*.78*	.02	-.00	.06	.20
Genuine – Persuasive	.45	.23	.16	.31	-.17
Composed – Contesting	.04	.24	-.06	.14	-.33
Optimistic – Pessimistic	-.07	-.11	-.35	.04	-.37
Abstract – Pragmatic	.17	.06	.10	*.77*	-.01

Note. Strongest loadings in italics

A comparison of the superfactors of both studies, using a target rotation procedure, yielded factorial agreement coefficients between .71 and .93. The low values of some factors were mainly due to secondary loadings; if the strongest factor loadings of a scale was scored as -1.0 for values lower than -.44, 0.0 for value between -.44 and .44, and as 1.0 for values above .44, the agreement of the factors was nearly perfect. The second factor showed the only discrepancy; the scales dealing with being persuasive and socially bold showed higher loadings on dominance in the first set and on extraversion in the second set. Given the generally high correspondence, we decided to use the same labels for the factors in both studies.

External Bias

Size of group differences. The effect size of cross-cultural differences for the predictor and criterion variables are reported in Table 4.9. All personality factors showed significant differences ($p < .01$) except for Stability. Medium effect sizes were found for Extraversion ($\eta^2 = .09$), Openness (.07), Agreeableness (.06), and Dominance (.05). Compared to the other groups, Blacks scored lowest on Extraversion, Openness, and Agreeableness, while Whites scored the lowest on dominance. There were no significant intergroup differences on the criterion scores. When these results are compared with the first study, it can be concluded that smaller cross-cultural differences were observed in the current study.

Table 4.9 Means of the Standardized Scores on the Scales of the Superfactors, Cognitive Skills and Criteria per Ethnic Group (Study 2)

Scale	Blacks	Coloreds	Whites	η^2
Extraversion	-0.24	0.41	0.35	0.09**
Dominance	0.18	-0.25	-0.31	0.05**
Agreeableness	-0.16	0.00	0.42	0.06**
Stability	0.04	0.02	-0.12	0.00
Openness	-0.20	0.25	0.37	0.07**
Cognitive Skill	0.13	-0.08	-0.28	0.03**
Criterion	0.05	-0.08	-0.09	0.01

Note. The last column gives the partial eta square values which indicate the proportion of variance accounted for by cultural group differences in the analysis of variance.

Multigroup structural equation modeling. The results of the regression analysis are reported in Table 4.10. The results of the fit test showed that the fit

did not significantly deteriorate when equality constraints on the regression coefficients. The introduction of equal intercepts showed a nonsignificant effect, $\Delta\chi^2(2) = 0.21$, $p = .90$. Introducing cross-cultural equality constraints on both the means of the predictors and their covariances yielded highly significant increments of the chi square statistic. Finally, holding the error term of the dependent variable identical across cultures did not lead to a significant increase of the fit. It can be concluded that there was unambiguous support for the cross-cultural identity of the slopes, intercepts, and the error variance of the dependent variable and equally unambiguous support for the heterogeneity of predictor means and covariances.

Table 4.10 Fit Statistics of the Multigroup Regression Analysis: Increments in Chi Square Statistics and Their Level of Significance (Study 2)

Model	$\Delta\chi^2$	Δdf	P
Structural weights	13.17	12	.36
Structural intercepts	0.21	2	.90
Structural means	170.0	12	.01
Structural covariances	211.30	42	.01
Structural residuals	7.75	2	.02

The estimated parameters of the model with an equal slope and intercept in the three ethnic groups are presented in Table 4.11. Positive and significant predictors were agreeableness and cognitive skill. The multiple correlations were .28 ($R^2 = .080$) for the Black group, .32 ($R^2 = .104$) for the Colored group, and .33 ($R^2 = .112$) for the White group. When compared to the results of the squared multiple correlation of the first study, we found a substantial overall increase in the multiple correlation.

Table 4.11 *Parameters of the Regression model with Equal Slopes and Intercepts in Each Cultural Group (Study 2)*

Predictor	b	β	*P*
Extraversion	0.47	-0.07	.28
Dominance	-0.70	0.04	.10
Agreeableness	1.07	0.11	.02
Stability	-0.21	-0.02	.63
Openness	0.08	0.01	.83
Cognitive Skill	2.68	0.25	.01
Slope	49.62		

Correction for unreliability and range restriction. The effects of reliability and range restriction are reported in Table 4.12. The selection process led to a reduction of the variance of on average 50%; the most affected variables were Agreeableness (76% reduction), Stability (63%) and Cognitive skill (50%). We find that after correction for unreliability and range restriction, the correlations were strongest and highly significant for agreeableness, stability, and cognitive skill. The selection ratios were very dissimilar for the three groups; the ratio was .03 for the Blacks, .08 for the Coloreds, and .25 for the Whites. Correcting for the unreliability of the criterion scores led to an increase of the multiple correlation of .28 to .36 for the Blacks, from .32 to .36 for the Coloreds, and .33 to .37 for the Whites. The change in mean scores of the cognitive abilities between the selected and total samples was larger than in the first study. Cognitive Skill and Stability continued to be good predictors. Overall, Table 4.12 provides additional evidence for the adequacy of the test adaptations of the second study.

Table 4.12 *Effect of Unreliability and Range Restriction on Correlation between Predictors and Criterion (Study 2)*

Predictor	*M and SD in total sample group (N = 16005)*	*M and SD in Selected Group (N = 597)*	Variance ratio	Correlation in selected group	Estimated correlation in total sample
Extraversion	*0 (0.24)*	*0.05 (0.22)*	0.61	-0.03	-0.05
Dominance	*0 (0.33)*	*-0.04 (0.34)*	0.68	0.03	0.04
Agreeableness	*0 (0.32)*	*-0.03 (0.20)*	0.24	0.10*	0.33**
Stability	*0 (0.20)*	*0.02 (0.14)*	0.37	0.10*	0.17**
Openness	*0 (0.32)*	*0.02 (0.24)*	0.54	-0.03	-0.07
Cognitive Skill	*54.74 (14.16)*	*65.98 (12.25)*	0.50	0.30**	0.51**

*$p < .05$. **$p < .01$.

The selection ratios were fairly consistent across the two studies for the Black and Colored groups, but went from .12 to .25 for the White group. We do not have a clear explanation of the finding. The effect may be due to cohort differences (better skilled White applicants in the second study).

Discussion

Study 2 employed adapted personality instruments in order to deal with some problems of the first study, notably as the low internal consistencies of the personality scales of the 15FQ+ and OPP. Strong evidence was found for the absence of internal bias; both the factors of the cognitive instruments and the superfactors of the personality questionnaires showed very good cross-cultural stability. We also addressed external bias and examined to what extent we complied with the technical criteria of fair prediction set forward by the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures. As to the first criterion, the use of an unbiased criterion, we replicated the findings of the first study: the police training criteria were unbiased. In terms of the second technical criterion (referring to the power to detect slope

and intercept differences), we reported more favourable reliability coefficients for the superfactors scales. The multigroup regression analysis supported the cross-cultural identity of the slopes, intercepts, and the reliability of the dependent variable, while predictor means and covariances were different. Multiple correlations between predictor and criterion showed increases across three cultural groups. The third criterion, dealing with the homogeneity of error variances, was also met in the current study. When we applied range restriction we reported substantial reduction in variance for most of the predictors, in line with the findings in the literature (e.g., Aguinis & Stone-Romero, 1997; Hunter & Hunter, 1984; Te Nijenhuis & Van der Flier, 2003).

The adapted personality questionnaires of the second study showed smaller cross-cultural differences than the original instruments, as employed in the first study. It is interesting to note that our findings are not in line with most literature, in which it is often reported that the removal of presumably biased items often does not reduce the size of cross-cultural differences (e.g., Abat, Colom, Robollo, & Escorial, 2004; Cole, Kawachi, Muller, & Berkman, 2000; Te Nijenhuis, Van der Flier, & Leeuwen, 2003). A possible reason for the discrepancy in findings may be the technique used to identify anomalous items. The studies reported in the literature used item differential functioning techniques to identify and remove biased items, whereas the current study used a content analysis, carried out by cultural experts, to change the wording and the contents of the items. It can be concluded that the adapted instruments complied better with the fairness guidelines than the versions used in the first study. Moreover, the improved internal consistencies of the personality questionnaires and the substantial increase in the squared multiple correlations of predictors provide support for the adequacy of the test adaptations.

General Discussion and Conclusion

We report the first analysis in South Africa in which both internal and external bias were studied concurrently. The cognitive instruments did not show internal bias; the scales of the original personality questionnaires showed low internal consistencies and some scales showed internal bias. However, the combination of the two questionnaires yielded five factors that were reliable and stable across ethnic groups. Adapted versions of the personality questionnaires, employed in the second study, yielded further improvements. External bias was addressed by examining to what extent the two studies complied with the technical standards put forward in the fourth edition of the Principles for the

Validation and use of Personnel Selection Procedures on predictive bias. Compliance with these technical standards is interpreted by us as adhering to legislative requirements of the Employment Equity Act in South Africa. The current test battery requires candidates from all cultural groups to meet the same standards. Alternative selection procedures such as quota hiring might lead to lowering the admission standards for some groups. We did not find evidence for differential prediction. The introduction of differential admission standards in our sample could lead to discrimination against the groups with higher scores. The findings are entirely in line with studies of validity generalization (Schmidt & Hunter, 1998).

Although the outcome may look very favorable in that it shows that the selection procedure meets the legal South African standards, we need to carefully analyze whether the selection procedure can be said to be fair. Fairness is a requirement of Section 8 of the Employment Equity Act. According to the Guidelines for the Validation and Use of Assessment Procedures for the Workplace (Society for Industrial and Organizational Psychology in South Africa, 2005), fairness is a social rather than a psychometric concept. Fairness is the total of all the variables that play a role or influence the final decision based on an assessment procedure. Equity plans play an important, legal role in fairness in South Africa. The South African Police Services are required to have equity plans in place which cater for equitable representation in all occupational categories and levels in the workforce. This could be read as a requirement for proportional representation of all ethnic groups and hence, for some type of quota hiring. In the South African Police Services equity targets for entry-level constables are set prior to the selection process and are aligned with the affirmative action plan of the organization. Post allocation of entry level positions in the South African Police Services is based on three criteria: demographic composition of the country, population make-up where entry-level applicants are recruited from, and organizational needs. In most cases, 80% of the posts are allocated to Blacks followed by Coloreds, Asians and Whites.

In the literature there is a well-known rule, called the four-fifth rule, which is also known as the 80 percent rule of thumb (Uniform Guidelines on Employee Selection Procedures, 1978). This rule stipulates that the selection ratio for a protected minority (in this case the majority Black group in South Africa) should not be less than four-fifth of the highest ratio for any of the other groups (i.e., Whites and Coloreds); lower ratios are taken to indicate adverse impact. According to the four-fifth rule, the test battery we used was biased. However, the

four-fifth rule does not have any formal status in South African employment legislation.

The question can be asked whether further refinements of the test battery would lead to bias-free measurement. In our view, it is difficult to eliminate both internal and external bias. An important source of potential bias is educational background. The educational quality of primary and secondary schools in South Africa varies greatly. The requirement that all applicants should have a "matric" certificate is an insufficient condition to reach homogeneity in the educational background of the applicants. Exposure to a poor education system in the past has an adverse impact on the majority Black groups, particularly in terms of their functional literacy in English. Undesirable influences of cross-ethnic differences in language knowledge on test performance is difficult to avoid. On the other hand, the influence of language knowledge may not be a problem in the assessment of external bias. Police training and examination is conducted in English, and in most cases formal statements and evidence in court have to be provided in English, which requires a good active and passive mastery of English. The biasing factor that may produce differences in test performance on the psychological instruments is also present in the criterion measure and constitutes an important element in the everyday functioning of the South African Police Services in which English is the language of communication of all official documents.

We found in this study that personality is less important than cognition and that only Extraversion, Stability, and Agreeableness added in validity to the cognitive measures. Cortina, Doherty, Schmitt, Kaufman, and Smith (1992) found similar results. They used a sample of police recruits and found that personality inventories did not add to the incremental validity. These authors argued that the questionnaires they used were less useful for police selection, because the tests were not developed specifically for the police. In the study by De Meijer, Born, Terlouw, and Van der Molen (2006), an adapted personality questionnaire was administered (the Police Personality Questionnaire) but the authors also reported low incremental validity for the personality measures. The lack of predictive validity for personality measures in police settings, also in the present study, may be attributed to poor predictor-criterion matching validation strategies. Lievens and Ones (2005) are of the opinion that most validation studies with personality scales are carried out with only limited time frames. Few studies have focused on the long-term predictive validity of personality scales. They found the predictive power of personality factors to increase over time. It can be argued that when police applicants are initially selected, certain personality traits are not yet very

salient but personality characteristics such as conscientious or openness become increasingly important. Stability to cope and ability to adapt to adversity of police work, especially in South Africa, will also be important in the police officer career later on. A meta-analysis by Drees, Ones, Cullen, Spellberg, and Viswesvaran (2003) on personality and police officer behaviors showed that personality scales are useful for predicting a broad range of differential criteria for police officers.
In conclusion, we found that the validation guidelines as set forward by SIOP can be applied adequately in a multicultural society such as South Africa and that results, obtained in western countries, generalize to South Africa; however, the present study shows that compliance with criteria of differential prediction does not yet guarantee fairness in a cultural context in which score differences on predictors are affected by confounding intergroup differences such as access to good education General

Author Note

We would like to acknowledge the feedback of Marise Ph. Born, Ype Poortinga, Murray Barrick and Stefan Mol on an earlier draft of the manuscript.

References

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo & I. Ebersohn. *South African Journal of Psychology, 32,* 58-61.

Abrahams, F., & Mauer, K. F. (1999a). The comparability of the constructs of the 16PF in the South African context. *Journal of Industrial Psychology, 25,* 53-59.

Abrahams, F., & Mauer, K. F. (1999b). Qualitative and statistical impact of home language on responses to the items of the Sixteen Personality Factor Questionnaire (16PF) in South African context. *South African Journal of Psychology, 29,* 76-86.

Abat, F. J., Colom, R., Robollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's advanced progressive matrices: evidence for bias. *Personality and Individual Differences, 36,* 1459-1470.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology, 82,*192-206.

Arbuckle, J. J. (2003). *Amos 5.0 user's guide.* Chicago, IL: Smallwaters.

American Psychological Association, American Educational Research, and National Council on Measurement in Education (Joint Committee). (1999). *Standards for educational and psychological test.* Washington, DC: American Educational Research Association.

Alonso, A., Viswesvaran, C., & Sanchez, J. I. (2001 April). *The mediating effects of task and contextual performance.* Poster session presented at the annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233-241.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). The FFM personality dimensions and job performance: Meta-analysis of meta-analysis. *International Journal of Selection and Assessment, 9,* 9-30.

Browne, M. W. (1979). The maximum likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology, 32,* 75-86.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications and programming.* Mahwah, NJ: Erlbaum.

Claassen, N. C. W. (1997). Culture differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47,* 297-307.

Chan, W., Ho, R. M., Leung, K., Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378-402.

Cortina, J. M., Doherty, M. L., Schmitt, N., Kaufman, G., & Smith, R. G. (1992). The Big Five Personality in the IPI and MMPI: Predictors of Police Performance. *Personnel Psychology,1,* 119-140.

Cole, S. R., Kawachi, I., Muller, S. J., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: experience for the New Haven EPESE study. *Journal of Clinical Epidemiology, 53,* 285-289.

Cudeck, R. (1982). Methods for estimating between-battery factors. *Multivariate Behavioral Research, 17,* 47-68.

Drees, S. A., Ones, D. S., Cullen, M. J., Spillberg, S. W., & Viswesvaran, V. (2003). *Personality assessment in police officers screening: Mandates and practice.* Paper presented at the 18th Annual Meeting of the Society of Industrial and Organizational Psychology. April 11, 2003. Orlando Florida.

De Meijer, L. A. L., Born, M. Ph., Terlouw, G., & Van der Molen, H. T. (2006, May). *Criterion-related validity of Dutch police officer selection and ethnic differences.* Paper presented in poster session at the 21st annual conference of the Society for Industrial and Organizational Psychology (SIOP), Dallas, TX.

Government Gazette, Republic of South Africa, Vol. 400, no. 19370. Cape Town, 19 October 1998.

Harkness, J. A. (2003) Questionnaire translation. In J. A. Harkness, F.J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York: Wiley.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance,18, 331-341*

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job performance relations. *Journal for Applied Psychology, 88,* 100-112.

Hough, L. M. (2001). I/Owes its advances to personality. In B. W. Roberts & R. T. Hogan (Eds.), *Personality psychology in the Workplace* (pp. 19-44) Washington, DC: American Psychological Association.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking towards the future-Remembering the past. *Annual Review of Psychology, 51,* 631-664.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152-194.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Department of Labor, Employment Services.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on mental testing (pp. 41-99).* New York: Plenum.

Houston, W. M., & Norvick, M. R. (1987). Race-based differential prediction in Air Force technical training programs. *Journal of Educational Measurements, 24,* 309-320.

Hurtz, G. M., & Donovan, J. J, (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869-879.

Lievens, F., Ones, D. S., & Dilchert, S. (2005, April). *Personality validity increases in medical school: A seven-year longitudinal investigation.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Meiring, D. (2000, June). *Revisiting the cross-cultural comparability of the 16 Personality Factor Inventory (16PF) in the South African context.* Paper presented at the Industrial Psychology Conference (incorporating the Psychometrics Conference), Pretoria, South Africa.

Meiring, D., Van de Vijver, F. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality measures in South Africa. *South African Journal of Industrial Psychology, 31,* 1–8.

Meiring, D., Van De Vijver, F. J. R., & Rothmann, S. R. (2006). Bias in an adapted version of the 15FQ+ questionnaire in South Africa. *South African Journal for Psychology. 36,340-356.*

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248-260.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376–398.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Uribina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101.

O'Conner, M. C. (1989). Aspects of differential performance by minorities on standardized test: Linguistic and sociocultural factors. In B. R. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp.129-181). Boston, MA: Kluwer Academic Publishers.

Ones, D. S., Viwesvaren, C. H., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance.* (in press)

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980) Validity generalization results for test used to predict job proficiency and training criteria in clerical occupations. *Journal of Applied Psychology, 65,* 373-407.

Psytech (2002a). *15FQ+ Fifteen Factor Questionnaire technical manual.* Retrieved December 2, 2005 from http:www.psytech.co.uk/downloads/manuals/15fq+%Bman.pdf.

Psytech (2002b). *Occupational personality profile, the technical manual.* Retrieved December 2, 2005 from http:www.psytech.co.uk/downloads/manuals/oppman.pdf

Saad, S., & Sackett, P. R. (2002). Investigating differential prediction by gender in employment-oriented personality measures. *Journal of Applied Psychology, 87,* 667-674

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist, 56*, 302-318.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist, 49,* 929-954.

Salgado, J. F. (1997). The five-factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82,* 30-43.

Salgado, J. F., Anderson, N., Moscos, S., Berutua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in European Community. *Journal of Applied Psychology, 88, 1068-1081.*

Salas, E., & Cannon-Browers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471-499.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162-173.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128-1137.

Schmidt, F. L., Pearlman, K. & Hunter, J. E. (1980). The validity and fairness of employment and educational stress for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705-724.

Society of Industrial and Organizational Psychology, Inc. (2003). *Principles for the Validation and Use of Personnel Selection Procedures* (4th ed.). College Park, MD: Author.

Society for Industrial and Organizational Psychology in South Africa (2005). *Guidelines for the Validation and Use of Assessment Procedures for the Workplace* (3rd ed.). Pretoria: Author.

Te Nijenhuis, J., & Van der Flier, H. (2000). Differential prediction of immigrant versus majority group training performance using cognitive and ability and personality measures. *International Journal of Selection and Assessment, 8,* 54-60.

Te Nijenhuis, J., Van de Flier, H., & Leeuwen, L. (2003). The use of a test for Neuroticism, Extraversion, and rigidity for Dutch immigrants job applicants. *Applied Psychology: An International review, 52,* 630-647.

Tredoux, N. (2004). *15FQ+ SA Reliability.* Retrieved March 1, 2004 from http://www.psytech.co.za/15FQPlus_reliabilities.htm

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika, 23,* 111-136.

Tyler, G. (2002). *A review of the 15FQ+ Personality Questionnaire.* Pulloxhill, UK: Psychometrics Limited.

Uniform Guidelines on Employee Selection Procedures. (1978). *Federal Register, 43,* 38290-38315.

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Wallice, T., & Birt, M. (2003). A comparison of native and non-native English-speaking groups' understanding of the vocabulary contained within the 16PF (SA92). *South African Journal of Psychology, 33,* 182–190.

Wightman, F.L. (1997). The threat to diversity in legal education: An empirical analysis of the consequences of abandoning race as a factor in law school decisions. *New York University Law Review, 72,* 1-53.

Wilson, K. M. (1981). Analyzing the long-term performance of minority and nonminority students: A tale of two studies. *Research in Higher Education,15,* 351-375.

Chapter Five
Personality Structure in South Africa: Commonalities of Three Comprehensive Measures of Personality Traits*

Abstract

Commonalities of three presumably comprehensive measures of personality (15 Factor Questionnaire, Occupational Personality Questionnaire, and Basic Trait Inventory) were examined in 1532 participants from eight cultural groups from across South Africa who had applied for a job at the South African Police Service. A combination of scales of three personality inventories yielded a new personality instrument with four stable factors with satisfactory psychometric properties. Equivalence of the four factors was found across all the cultural groups. Three of the Big Five factors were found in this multicultural South African population, namely Emotional Stability, Extraversion and Openness to Experience. The Extraversion factor was split into a Dominance and a Sociability factor. Although many items of the new instrument showed significant item bias, the cross-cultural differences in scale scores were not influenced by the removal of the biased items; these differences were not affected by social desirability, either.

*Meiring, D., Van de Vijver, F., De Bruin, G.P. & Rothmann, S. (2006). *Personality Structure in South Africa : Commonalities of Three Comprehensive Measures of Personality Traits.* (Manuscript submitted for publication).

There is much cross-cultural research on the structure of personality. Evidence for the pancultural identity of personality structure has been found for both the Five-Factor Model (McCrae & Allik, 2002; McCrae & Costa, 1997) and Eysenck's Personality Model (Barrett, Petrides, Eysenck, & Eysenck, 1998; Van Hemert, Van de Vijver, Poortinga, & Georgas, 2002). The focus on using the same instrument in various cultures and on finding cross-cultural commonalities ("etic aspects") in personality structure has detracted attention from identifying culturally specific aspects of personality ("emic aspects"). The question can be raised whether the universal structure that is supported by the above studies provides a comprehensive description of personality. There are two ways to address this issue; existing or adapted instruments can be examined or new questionnaires can be developed. The first approach was adopted in this study. The current study examines personality structure in various cultural groups in South Africa. Commonalities were investigated of three separate comprehensive personality measures, 15 Factor Questionnaire (15FQ+; Psytech, 2002a), Occupational Personality Profile (OPP; Psytech, 2002b), and the Basic Trait Inventory (BTI; Taylor & De Bruin, 2005). The first two are imported instruments, while the BTI (based on the Five-Factor Model) was specifically designed for cross-cultural use in South Africa with careful consideration of item content, structure, and presentation. We combined elements of the 15FQ+, OPP, and the BTI in order to form a new instrument.

When personality measures are applied cross-culturally, issues of measurement bias and equivalence become important (Church, 2001). Thus, items developed in a particular country might better represent the construct in that country than in other countries. The underlying dimensions of personality may not be different across cultures, but expressions of personality are more likely to differ (Church & Katigbak, 1988). Imported instruments are often adapted extensively, which introduces the need to examine the similarity of the psychological meaning of measures used across different language versions (Hambleton & Kanjee, 1995). The cross-cultural suitability of the new instrument is studied here by examining three kinds of bias (Van de Vijver & Leung, 1997). The first is construct bias, addressing the question of whether the instrument measures the same personality dimensions across all cultural groups in South Africa. We also addressed item bias (differential item functioning). The size of cross-cultural differences was compared before and after presumably biased items were removed. Finally, we addressed method bias (a generic term for all method-related aspects of an instrument that could give rise to cross-cultural differences);

more specifically, we examined the influence of social desirability on the cross-cultural differences observed with the instrument.

Personality Structure across Cultures

According to Church (2000), research on imported personality tests generally investigates the universality of personality dimensions or the nomological networks of personality constructs across cultures or compares trait levels across cultures. In contemporary personality psychology the model that has received the most attention is the Five-Factor Model (see De Raad & Perugini, 2002; McCrae & John 1992; Paunonen & Ashton, 1998). The Five-Factor Model is a hierarchical model of trait structure, in which relatively narrow and specific traits are organized in five broad factors: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. Replicability of the Five-Factor Model has been studied extensively across cultures (see Cheung et al., 2001; McCrae & Allik, 2002; McCrae & Costa, 1997; Paunonen, Zeidner, Enggvik, Oosterveld, & Maliphant, 2000). The NEO PI-R (NEO Five–Factor Inventory) of Costa and McCrea (1992) is the most widely used in this respect and dozens of translations and adaptations have been made. In general, the five dimensions of the Five-Factor Model that were identified originally in the USA are also found elsewhere including non-Western societies such as China, Korea, Russia, Israel, Philippines, and Japan (for a review, see McCrea & Costa, 1997; McCrea, Costa, Del-Pilar, Rolland, & Parker, 1998).

Recently, De Raad and Peabody (2005) contested the stability of the Five-Factor Model across cultures. The researchers re-examined the five-factor analytic solutions of six European studies and found that the five-factor results did not always support the Five-Factor Model. Their re-analyses of existing data suggested that a three-factor model (i.e., Extraversion, Agreeableness, and Conscientiousness) provides a more appropriate fit; yet, two of the three factors often split up in lower-order factors, in which case Neuroticism and Openness emerged. The authors claim that these Big Three are universal, while the universality of the full Big Five may be restricted to American English and German speakers.

Another three-factor model of personality structure (the "Giant Three") was proposed by Eysenck (1970, 1992). The three factors are Extraversion, Neuroticism, and Psychoticism. The Eysenck Personality Questionnaire (EPQ) contains scales for each factor, alongside a Lie scale. Cross-cultural replications of Eysenck three-factor model have been carried out in 34 countries (Barrett et al., 1998). The similarity of factors found in these countries with the U.K. factors

was high for Extraversion and Neuroticism, while the cross-cultural stability of the Psychoticism and Lie Scales was sometimes lower, though usually still acceptable.

Studies of personality or studies in which emic and etic aspects are combined are not common. The Chinese Personality Assessment Inventory (CPAI) is an example of an indigenous instrument that was designed to measure traits considered to be important to the Chinese (Cheung et al., 1996). In a subsequent study, Cheung and colleagues compared the CPAI with the NEO-PI-R. They found in a factor analysis of the combined data that three of the CPAI personality factors (Dependability, Social Potency, and Individualism) converged with Neuroticism, Extraversion, and Agreeableness of the NEO-PI-R (Cheung et al., 2001). The Interpersonal Relatedness factor of the CPAI did not load on any of the NEO-PI-R facets, and none of the CPAI factors loaded on the Openness facet of the NEO-PI-R. Despite the universality of the Five-Factor Model, which has also been found in several Chinese samples, the unique Interpersonal Relatedness factors suggest that Chinese people cut the social-perceptual world in a slightly different way. Lin and Church (2004), investigating the replication of the Interpersonal Relatedness dimensions of the CPAI in Chinese American and European American samples, found good replication, suggesting that this dimension is not culture specific. Still, the authors indicate that this dimension is more characteristic of individuals who retain or identify with the Chinese culture. The CPAI was revised in 2000 to include a set of openness scales for normal personality assessment. The openness scales are a combination of indigenous aspects, such as novelty and diversity and etic aspects from the Five-Factor Model such as Aesthetics. The new instrument showed a very stable structure across different cultural groups (Mainland China, Taiwan, Korea, Japan as well as Asian American and Caucasian American students). It is interesting to note that the Interpersonal Relatedness Factor was also stable across cultures (Cheung, 2006).

Personality Structure in Sub-Saharan Africa

Cross-cultural research on personality questionnaires in sub-Saharan Africa has been limited. In a recent study by Piedmont, Bain, McCrea, and Costa (2002), the applicability of the five-factor mode in a sub-Saharan culture was explored. The purpose of the study was to evaluate the utility of the Five-Factor Model in Shona, a native language of Zimbabwe. It was found that translating the NEO PI-R into Shona was difficult. Some facets measured by the NEO PI-R have no counterpart in the Shona culture. For example, as in some other collectivist

cultures altruism is an unknown concept. Shona will not understand why someone would donate time, energy, and money to an unknown person without receiving anything in return. Most sub-Saharan studies were carried out in South Africa. A study by Heuchert, Parker, Strumf, and Myburg (2000), using the NEO-PI-R with college students, showed that the fit of the model was only tenuous for both Black and White students. In an unpublished thesis, Horn (2000) examined a Xhosa translation (Xhosa is a Bantu language spoken in South Africa) of the NEO PI-R. The author reported that translation was difficult, in part because the Xhosa language has a restricted vocabulary with regard to personality descriptives. Some scales of the NEO PI-R showed low internal consistency estimates.

Taylor (2000) conducted a construct comparability study of the NEO PI-R for Black and White employees in a work setting in South Africa. The NEO PI-R did not work as well for Blacks as it did for Whites; in particular, the Openness factor was not found in the Black sample. According to De Bruin (2005), attempts to isolate the Big Five factors among South Africans have yielded mixed (but mostly disappointing) results when common instruments are employed. These results may be attributed in part to the cultural inappropriateness of some of the items of imported questionnaires and the complexity of the item wordings. Many people in South Africa do not have English as their first language, which means that some of the items of imported questionnaires may be poorly understood.

In response to these results, Taylor and de Bruin (2005) developed the Basic Traits Inventory as an English-language measure of the Five-Factor Model to be used across all ethnic groups in South Africa. The BTI has a hierarchical structure similar to that of the NEO PI-R (Costa & McCrae, 1992), with five broad traits on the highest level and 24 facets that serve as indicators of the factors on a lower level. The facets were selected on the basis of a review of factor-analytic studies of Big Five instruments. Attempts were made to select facets so that the behaviors associated with each of the five factors are comprehensively covered, while avoiding facets with salient loadings on more than one factor. Taylor and de Bruin attempted to maximize the cross-cultural suitability of the BTI by screening all items for appropriateness with regard to content and comprehensibility (they only used short items), presenting the items in content blocks, avoiding reverse scored items, and clearly labeling the response categories of the five-point Likert-type response scale. Results showed that the expected Big Five structure was found and that it replicated well across different cultural groups in South Africa. Ramsay, Taylor, De Bruin, and Meiring (2005) subjected the BTI responses of Nguni (Zulu, Xhosa, Ndebele, and Swazi), Sotho (Tswana and South-Sotho) and Pedi speakers to a multiple-group confirmatory factor analysis. The participants

had a completed secondary-school education and were proficient in English. The authors found a very good fit between the postulated five-factor model and the observed data in each group. These results provide strong support for the construct validity of the BTI with Black South African participants. Other studies have compared the structure of the BTI for White and Black participants and showed that the five factors replicate very well in these groups, with Tucker coefficients of congruence above .90 for all factors (De Bruin, Taylor, & Schepers, 2005).

Other personality tests have also been investigated for cross-cultural appropriateness in South Africa, such as the South African Personality Questionnaire (SAPQ) and 16PF (South African 1992 Version; SA 92) (Abrahams, 1996; Boeyens & Taylor, 1991; Spence, 1982; Tack, 1998). The main findings indicated the absence of construct equivalence across the different cultural groups; moreover, individuals whose first language was not English experienced problems with these questionnaires.

Two recent studies investigated the cross-cultural adequacy of the Fifteen Factor Questionnaire (15FQ+) in South Africa (Meiring, Van de Vijver, & Rothmann, 2006; Meiring, Van de Vijver, Rothmann, & Barrick, 2006). Both studies showed a poor structural equivalence in various ethnic groups for both the original 15FQ+ and an adapted version in which various items used a simpler language than in the original version. The suitability of the instrument was further compromised by the low internal consistencies of some scales.

The current study set out to examine South-African personality structure by employing three presumably comprehensive personality measures. To the best of our knowledge, the current study is the most comprehensive examination of personality in South Africa so far by addressing cross-battery variance of three comprehensive personality measures and by including various language groups. South Africa is remarkably heterogeneous in terms of cultures. Therefore, a test of commonalities of three presumably comprehensive measures provides an interesting test of the universality of personality structures.

Method

Participants

The sample included 1532 participants across South Africa who had applied to become a police official in the South African Police Service. The sample consisted of Blacks (n = 1277), Whites (n = 27), and Coloureds (n = 196) (32 missing values); 72% (n = 1107) were male and 27% (n = 410) were female (15 missing values). The Black group consisted of the following cultural groups:

Sepedi (n = 212), SeSotho (n = 110), Setswana (n = 250), Swati (n = 115), Tsonga (n = 103), Xhosa (n = 231), and Zulu (n = 287); the Venda and the Ndebele groups were excluded due to their small sample sizes. The mean age of the participants was 25 years (SD = 4.80). The entry level qualification for the police is Grade 12; 78% of the sample group had this qualification, 18% had a degree or diploma and 1% had a postgraduate qualification.

Instruments

The 15FQ+ is a 200-item personality test with a trichotomous response scale, developed by Psytech International (Psytech, 2002a). The instrument attempts to provide a relatively short, yet robust measure of Cattell's primary personality factors. The scales of the 15FQ+ measures are described in Table 5.1. For the purpose of this study the Intellectance scale was excluded as it relates to some form of reasoning ability behavior. The questionnaire also has additional measures of social desirability and infrequency. The manual reports satisfactory internal consistencies of the scales, with mean values of above .70 for most of the scales (Psytech, 2002a). An adapted version of the 15FQ+ was used for this study (see Meiring et al., 2006, for a full discussion of the adaptation). The internal consistencies means per culture group as found in the current study reported in Table 5.1; as can be seen there, the problem of the low internal consistencies found before (Meiring, Van de Vijver, Rothmann, & Barrick, 2006; Meiring, Van de Vijver, & Rothmann, 2006) was replicated here. Taylor (2002) reports evidence supporting the construct validity of the 15FQ+ as measured by correlations with other personality measures (16PF, 16PF5, and Five-Factor Model).

The Occupational Personality Profile (OPP), also developed by Psytech International (2002b), is a personality measure for use in industrial and organizational settings. The inventory consists of 98 items and uses a five point-Likert response format ranging from *strongly agree* (1) to *strongly disagree* (5) throughout the questionnaire. The OPP measures nine different personality dimensions. Reliability coefficients above .60 have been reported for all constructs of the OPP. For this study the internal consistencies means per culture group are reported in Table 5.1; again, the internal consistencies were not high. The manual of the OPP reports favorable results for both the construct and predictive validity of the OPP (Psytech, 2002b). The adaptation process, which was similar to the adaptation of the 15FQ+, focused on increasing the cultural and linguistic contents of the items for the South African context.

The Basic Traits Inventory (BTI; Taylor & De Bruin, 2005) is a 193-item inventory that has a five-point Likert-type response scale that is used throughout

the questionnaire, ranging from *strongly disagree* to *strongly agree*. The BTI was developed as a personality instrument to be used in the multicultural South African context in order to assess the Big Five factors (Extraversion, Neuroticism, Conscientiousness, Openness to Experience, Agreeableness) of personality. Each of the five factors is made up of four to five facets, which measure different aspects of their respective factor. For each of the five factors, the Cronbach alpha coefficients were satisfactory across the English, Afrikaans, Sotho, and Nguni groups (Taylor & De Bruin, 2005). The mean alpha coefficients across all groups were .87 for Extraversion, .93 for Neuroticism, .87 for Openness to Experience, .93 for Conscientiousness, and .87 for Agreeableness. Mean reliabilities for the BTI across the different culture groups are also reported in Table 5.1. These values pointed to a good internal consistency of the BTI in all culture groups.

Table 5.1 *Values of Cronbach's Alpha for the 15FQ+ OPP and BTI per Cultural Group*

	Cultural Group							
Questionnaire/ Scale	Afrikaans	Sepedi	Sesotho	Setswana	Swati	Tsonga	Xhosa	Zulu
15FQ+								
Cool Reserved – Outgoing	.61	.46	,40	.52	.46	.40	.52	.46
Affected by Feelings - Emotionally Stable	.59	.47	.31	.43	.37	.38	.53	.41
Accommodating – Dominant	.56	.41	.35	.36	.52	.41	.33	.44
Sober Serious – Enthusiastic	.68	.67	.65	.67	.72	.62	.58	.66
Expedient – Conscientious	.31	.10	.23	.07	.30	.10	.41	.10
Retiring – Socially Bold	.71	,71	.48	.68	.61	.53	.60	.64
Tough Minded – Tender Minded	.71	.56	.57	.63	.58	.50	.41	.55
Trusting – Suspicious	.58	.63	.61	.60	.51	.52	.58	.53
Practical – Abstract	.45	.16	.35	.14	.01	.21	.16	.30
Forthright – Discreet	.55	.10	.12	.37	.41	.20	.31	.11
Self-assured – Apprehensive	.46	.28	.37	.37	.41	.31	.32	.26
Conventional – Radical	.50	.33	.40	.29	.06	.08	.34	.32
Group – Orientated – Self-Sufficient	.62	.61	.64	.59	.63	.56	.48	.54
Undisciplined – Self-Disciplined	.36	.04	.06	.10	.20	.25	.00	.36

Relaxed – Tense Driven	.57	.02	.03	.11	.19	.02.	.12	.15
OPP								
Accommodating – Assertive	.59	.44	.44	.52	.54	.44	.47	.51
Detail-Conscious – Flexible	.72	.77	.73	.70	.73	.48	.84	.83
Cynical – Trusting	.75	.76	.66	.75	.63	.75	.64	.75
Emotional – Phlegmatic	.77	.71	.76	.70	.71	.73	.66	.77
Reserved – Gregarious	.77	.72	.63	.71	.76	.65	.60	.71
Genuine – Persuasive	.64	.74	.68	.74	.75	.68	.62	.69
Composed – Contesting	.70	.55	.63	.60	.56	.69	.62	60
Optimistic – Pessimistic	.65	.70	.69	.60	.51	.62	.60	.61
Abstract – Pragmatic	.76.	.81	.82	.78	.80	.70	.75	.81
BTI								
Extraversion								
Ascendance	.81	.80	.76	.80	.81	.75	.74	.76
Liveliness	.70	.58	.68	.73	.70	.56	.61	.56
Positive emotions	.76	.71	.64	.75	.64	.66	.70	.66
Gregariousness	.79	.84	.71	.82	.83	.80	.76	.81
Excitement-seeking	.77	.68	.67	.72	.62	.61	.69	.68

Table 5.1 (continued) *Values of Cronbach's Alpha for the 15FQ+ OPP and BTI per Cultural Group*

	Cultural Group							
Questionnaire/ Scale	Afrikaans	Sepedi	Sesotho	Setswana	Swati	Tsonga	Xhosa	Zulu
Neuroticism								
Depression	.83	.79	.74	.74	.68	.63	.81	.76
Self-consciousness	.79	.75	.64	.68	.51	.64	.70	.65
Anxiety	.88	.86	.70	.76	.68	.65	.77	.79
Effort	.86	.77	.75	.78	.64	.71	.79	.79
Conscientiousness								
Order	.90	.81	.80	.80	.74	.81	.86	.84
Dutifulness	.80	.80	.77	.73	.63	.70	.82	.81
Prudence	.69	.74	.70	.67	.68	.44	.83	.78
Self-discipline	.87	.80	.76	.73	.74	.72	.75	.73
Openness to Experiences								
Aesthetics	.83	.79	.76	.78	.72	.76	.73	.77
Ideas	.68	.75	.72	.76	.75	.72	.70	.72
Actions	.72	.73	.67	.70	.76	.61	.76	.75
Values	.38	.39	.45	.34	.23	.31	.44	.47

Imagination	.81	.79	.80	.80	.73	.61	.79	.68
Agreeableness								
Straightforwardness	.65	.71	.64	.60	.59	.53	.70	.66
Compliance	.81	.67	.66	.68	.66	.57	.78	.69
Prosocial tendencies	.83	.78	.81	.80	.78	.77	.86	.80
Modesty	.76	.69	.74	.73	.63	.55	.59	.67
Tendermindedness	.83	.82	.81	.73	.76	.82	.80	.72

Procedure

The instruments were administered using a standardized procedure by previously trained personnel of Psychological Services of the South African Police Service. The test session lasted 3 hours with a break of 15 minutes. Computer-readable answer sheets were employed for all tests.

Statistical Analysis

The analysis involved five steps. *Step 1* involved an interbattery factor analysis of the three instruments at item level. We wanted to extract the factors common to three personality instruments without finding battery or method-specific factors (e.g., due to differential skewness) (Schepers, 2003). Interbattery factor analysis has been proposed as a technique that can overcome battery- or method-specific factors (Tucker, 1958; see also Browne, 1979). Interbattery factor analysis allows a clearer empirical understanding of what is and what is not shared by measures by analyzing only the cross-battery correlations of the variables. We employed this analysis (Browne, 1979) to identify those factors that are common to the 15FQ+, OPP, and BTI variables. The technique identifies dimensions that are common to the underlying factor structure of the three instruments. This method is an attractive alternative to exploratory factor analysis due to its ability to separate within-battery and cross-battery variability (Cudeck, 1982). With the exclusive focus on cross-battery variability, it is also possible that some scales within a battery will not load on any of the interbattery factors that are obtained.

Step 2 involved the psychometric analysis of the new factors. It included item screening by means of classical item analysis and estimating internal consistencies for each of the new factors.

Step 3 entailed the evaluation of the internal bias (structural equivalence) of the new factor structure. Factors obtained in each of the eight language cultural groups are compared to the factors obtained in the pooled group, in which all data are combined (Muthén, 1991, 1994). Factors in the cultural groups were rotated to the factors of the pooled solution. The agreement was evaluated by means of a factor congruence coefficient, Tucker's phi (Chan, Ho, Leung, Cha, & Yung, 1999; Van de Vijver & Leung, 1997). Values above .90 are taken to point to essential agreement and values above .95 to very high agreement.

Step 4 entailed an analysis of item bias. Item level analysis (item bias analysis) was performed by using analysis of variance (ANOVA) for the three personality questionnaires (yielding interval-level scores). The assumption is that an item is unbiased if persons from different cultures with an equal standing on

the construct underlying the instrument have the same expected score on the item (Van de Vijver & Leung, 1997). The independent variables in the ANOVA were ethnicity (eight levels) and score level (four levels). The latter was based on the sum of the standardized item scores of all items belonging to a scale; the standardization was needed as each of the three personality inventories uses its own number of response alternatives. The item score was the dependent variable, while culture and score levels were the independent variables. A significant main effect of ethnic group was taken to point to uniform bias and a significant interaction of score level and ethnic group pointed to non-uniform bias (Van de Vijver & Leung, 1997). The dependent variables were the item scores. We were more interested in the size of the bias than in its significance, given the large sample size. Therefore, we used Cohen's (1988) criteria of effect sizes. We use three threshold criteria of .01 as very small, .035 (halfway between small effect and medium effect), and .06 as a medium-effect size. The latter size was chosen as it can be considered to be large enough to be practically important.

In *Step 5* the role of bias (differential item functioning) in explaining cross-cultural differences for item bias was established by comparing the size of cross-cultural differences in mean scores before and after correction for item bias; an analysis of variance was carried out on the mean scale scores based on the full item set. The effect size found in this analysis was compared to the size found in the analysis of variance based on the mean scale scores from which all the presumably biased items are removed. A final analysis involved the influence of social desirability. Cross-cultural differences in mean scale scores were compared before and after correction for social desirability in an analysis of covariance.

Results

Interbattery factor analysis. An interbattery factor analysis using the non-iterative PACE algorithm of the MultiPace software (Browne & Tateneni, 1997) was carried out in the first step. Three-, four-, and five-factor solutions were obtained and rotated according to the oblique Direct Quartimin criterion. Inspection of the cross-battery residual matrix showed that four factors did an adequate job in accounting for the commonalities of the three batteries. Furthermore, the fifth factor did not show any loadings with an absolute value larger than .30, suggesting that this factor was poorly defined. The four-factor solution produced four well-defined and interpretable factors. The factor pattern matrix is presented in Table 5.2.

The first factor appears to represent *Emotional Stability* and each of the three batteries appears to measure this factor rather well. Four 15FQ+ scales that typically define Cattell's second-order Anxiety factor (Krug & Johns, 1986) loaded on this factor: Affected by Feelings–Emotionally Stable (.50), Trusting–Suspicious (-.30), Self-assured–Apprehensive (-.51), and Relaxed–Tense Driven (-.41). Three OPP scales that appear directly related to emotional stability loaded on this factor, namely Cynical–Trusting (.38), Emotional–Phlegmatic (.67), and Optimistic–Pessimistic (.43). Finally, four BTI facets that are designed to measure a Neuroticism factor, loaded on this factor, namely Affective instability (-.41), Depression (-.49), Self-Consciousness (-.49), and Anxiety (-.52).

The second factor appears to measure *Social Dominance*. Two 15FQ+ scales loaded this factor, namely Accommodating–Dominant (.39) and Retiring–Socially Bold (.43). These two scales are normally associated with Cattell's second order Extraversion factor, but Accommodating–Dominant also commonly loads on an Independence factor. The three OPP scales that loaded on this factor, namely Accommodating – Assertive (.47), Genuine–Persuasive (-.65) and Composed–Contesting (-.37) also appear to measure aspects of social dominance. Two BTI facets that are designed to measure Extraversion, namely Ascendance (.58) and Liveliness (.38), and one designed to measure Conscientiousness, namely Effort (.33), loaded on this factor.

The third factor appears to measure *Sociability*. The two 15FQ+ scales that loaded on this factor, namely Sober serious–Enthusiastic (.61) and Group-Orientated–Self-Sufficient (-.40), typically define Cattell's second-order Extraversion factor. The OPP scale that loaded on this factor, Reserved–Gregariousness (.67,) appears to be directly related to sociability. Three BTI facets that are designed to measure Extraversion, namely Positive emotions (.30), Gregariousness (.71), and Excitement-seeking (.38), loaded on this factor.

The fourth factor appears to measure aspects of an *Openness to Experience* factor. The 15FQ+ scale that loaded on this factor, namely Toughminded–Tenderminded (.45), usually loads on Cattell's second-order Tough-mindedness factor. The OPP scale that loaded on this factor, namely Abstract–Pragmatic (-.60) appears to be related to Openness to Experience. One BTI facet designed to measure the Openness factor, namely Aesthetics (.66), had a salient loading.

Each of the three instruments contained subscales that did not show a high loading on any factor. More specifically, several facets of the BTI that are designed to measure Agreeableness and Conscientiousness failed to load on any of the four interbattery factors. Also, one OPP scale (Detail Conscious–Flexible)

and six scales of the 15FQ+ (Cool Reserved–Outgoing, Expedient–Conscientious, Practical–Abstract, Conventional–Radical, Undisciplined–Self-Disciplined) failed to load on any of the four interbattery factors. This result shows that these facets and scales reflect variance not shared with the other batteries (which is a prerequisite to define a factor in interbattery factor analysis).

Table 5.2 *Rotated Interbattery Factor Solution of 15FQ+,OPP, and BTI Scales*

Scales	Emotional Stability	Dominance	Sociability	Openness to experience
		15 Factor Questionnaire		
Cool Reserved - Outgoing	.05	-.01	.29	.18
Affected by Feelings – Emotionally Stable	*.50*	.03	-.02	-.04
Accommodating – Dominant	.10	*.39*	.10	-.14
Sober serious – Enthusiastic	.00	.02	*.61*	-.09
Expedient – Conscientious	.08	.26	-.10	.12
Retiring – Socially Bold	.27	*.43*	.17	-.02
Tough Minded - Tender Minded	-.12	.04	-.03	*.45*
Trusting – Suspicious	*-.30*	.08	-.21	-.13
Practical – Abstract	-.25	.22	.04	.09
Forthright – Discreet	.04	-.07	-.07	.15
Self-assured – Apprehensive	*-.51*	-.10	.09	.07
Conventional – Radical	-.02	-.02	.06	-.08
Group - Orientated - Self-Sufficient	-.02	.07	*-.40*	-.01
Undisciplined - Self-Disciplined	.07	-.05	.06	.04
Relaxed - Tense Driven	*-.41*	.16	-.06	.01
		Occupational Personality Questionnaire		
Accommodating – Assertive	.08	*.47*	.02	-.07
Detail-Conscious – Flexible	-.06	-.02	-.02	*-.09*

Cynical – Trusting	*.38*	-.05	.05	.09
Emotional – Phlegmatic	*.67*	.17	-.01	-.08
Reserved – Gregarious	.15	.02	*.67*	-.07
Genuine – Persuasive	.02	*-.65*	-.23	-.07
Composed – Contesting	.25	*-.37*	.00	-.07
Optimistic – Pessimistic	*.43*	-.07	-.03	-.06
Abstract – Pragmatic	.08	*-.30*	-.07	*-.60*
	Basic Trait Inventory			
Extraversion				
Ascendance	.05	*.58*	.09	.13
Liveliness	.06	*.38*	.26	.04
Positive emotions	.09	.09	.30	.20
Gregariousness	.04	.15	*.71*	-.01
Excitement-seeking	-.13	.16	*.38*	-.16
Neuroticism				
Affective instability	*-.41*	-.02	-.08	-.11
Depression	*-.49*	-.09	-.07	-.03
Self-consciousness	*-.49*	-.15	-.01	-.06
Anxiety	*-.52*	-.10	-.02	.00
Conscientiousness				
Effort	.12	*.33*	-.02	.13
Order	.20	.16	.01	.14
Dutifulness	.13	.16	.03	.15
Prudence	.07	.25	-.09	.23
Self-discipline	.23	.18	.06	.11
Openness to Experience				
Aesthetics	-.13	.06	.11	*.66*

Ideas	.00	.26	.25	.28
Actions	-.04	.20	.29	.29
Values	-.07	.15	.04	.05
Imagination	.00	.28	.12	*.24*
Agreeableness				
Straightforwardness	.08	.15	.18	.20
Compliance	.06	-.11	.24	.26
Prosocial tendencies	.09	.25	.05	*.33*
Modesty	.01	-.13	.06	.00
Tendermindedness	.09	.00	.13	.19

Scales with loadings (absolute value) larger than .30 are printed in italics.

Item analysis and internal consistencies. The second step examined the internal consistencies of the five factors. Table 5.3 reports the alpha reliability coefficients for the four factors. The reliability coefficients were satisfactory across all eight culture groups. These reliabilities are much higher than the values found in incomparable cultural groups for the original scales of the 15FQ+ and OPP as indicated in Table 5.1. The problem of low values of the internal consistency estimates has been largely solved with the new constituted factors from the interbattery factor analysis.

Table 5.3 *Values of Cronbach's Alpha of the Scales, Based on the Interbattery Factor Analysis, in the Eight Language Groups*

Scale	Afrikaans	Sepedi	Sesotho	SetSwana	Swati	Tsonga	Xhosa	Zulu
Emotional Stability	.81	.76	.77	.76	.70	.60	.73	.75
Dominance	.88	.87	.87	.87	.90	.89	.86	.89
Sociability	.85	.83	.86	.79	.87	.89	.86	.87
Openness to Experience	.85	.75	.82	.81	.83	.78	.78	.81

Note. Length of scales: Emotional Stability: 113 items; Dominance: 77 items; Sociability: 55 Items; Openness to Experiences: 29 items.

Structural equivalence. The third step involved the evaluation of internal bias. Table 5.4 shows the agreement of the factors derived from the pooled data with the factors in the eight culture groups. Values of Tucker's phi higher than .90 were found for all the culture groups except for the Setswana group on the Sociability factor, which shows a value of Tucker's phi of .88. This provided a strong indication of the structural equivalence of the four factors underlying the performance of all the different groups distinguished.

Table 5.4 *Values of Tucker's Phi (Factorial Agreement) of the Factors of the Interbattery Factor Analysis for Eight Cultural Groups*

Scale	Afrikaans	Sepedi	Sesothoh	SetSwana	Swati	Tsonga	Xhosa	Zulu
Emotional Stability	.98	.98	.95	.97	.95	.95	.97	.97
Dominance	.96	.98	.95	.98	.96	.96	.97	.98
Sociability	.97	.98	.88	.98	.97	.93	.96	.99
Openness to Experience	.96	.99	.97	.98	.97	.95	.97	.99

Item bias. As can be seen in Table 5.5, the majority of the items in each of the four scales showed statistically significant bias, but very few items showed medium or large effect sizes.

Table 5.5 *Item Bias Analyses: Percentage of items with Significant Effect Sizes for Uniform and Non-Uniform Bias per Scale*

	Significant effect sizes							
	< .010		Between .010 and .035		Between .035 and .060		> .060	
Scales	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform
Emotional Stability	27.1	7.9	7.9	0	4.5	0	3.4	0
Dominance	10.0	.8	0.8	0	0.8	0	0.8	0
Sociability	2.8	.6	2.2	0	1.1	0	0.0	0
Openness	1.5	.6	0.3	0	1.5	0	0.6	0

Note. Numbers in represent proportions of items of a scale with a particular effect size; for example; 27.1% of the items of the Emotional Stability Scale showed effect size of at most .01. Items that showed significant uniform and non-uniform bias are counted as separate items in the Table.

The role of item bias and social desirability in explaining cross-cultural differences. In order to inspect the impact of item bias on cross-cultural differences, the size of these differences was computed before and after the elimination of biased items. An item was eliminated if its eta square value for uniform or non-uniform bias was at least .02. This relatively small value was arbitrarily chosen because of the overall low level of the effect sizes. One-way analyses of variance were carried out with language group as independent variable and scale scores (sum scores on the items pertaining to the scale) as dependent variables. In a second step the procedure was repeated, but now all biased items were excluded from the computation of scale scores. The extent of the cross-cultural differences was evaluated as the effect size (eta square) of the culture component. The mean effect size was .041 before the removal of biased items and .039 after the biased items were removed. It can be concluded that the correction for biased items did not affect the size of the cross-cultural differences observed.

The influence of social desirability was investigated in analyses of covariance in which ethnic group was the independent variable, social desirability the covariate, while the four scale scores were the dependent variables. Results are presented in Table 5.6. Social desirability reduced the average effect size from .023 to .019. The influence of social desirability was small both before and after correction.

Table 5.6 *Analysis of Covariance with Ethnic Group as Independent Variables, Social Desirability as the Covariate, and the Five Factor Scores as Dependent Variables*

Scale	Effect size before correction	Effect size after correction
Emotional Stability	.016	.007
Dominance	.031	.029
Sociability	.030	.032
Openness to Experience	.016	.009

Discussion

Personality structure in South Africa was studied by examining commonalities of three personality questionnaires (15 Factor Questionnaire, Occupational Personality Questionnaire, and Basic Trait Inventory). The instruments were administered to a group of 1532 job applicants at the South African Police Service. An interbattery factor analysis showed that a combination of scales of three personality inventories yielded a new personality instrument with four stable factors with satisfactory internal consistencies. Structural equivalence of the four factors was found across eight cultural groups. Three of the Big Five were found in this multicultural South African sample namely Emotional Stability, Extraversion, and Openness to Experience. The Extraversion factor was split into a Dominance factor and a Sociability factor. Although many items of the new instrument showed some statistically significant bias, the cross-cultural differences in scale scores were hardly influenced by the removal of the biased items. In addition, we found a very small effect of social desirability on cross-cultural differences; the effect sizes were small both before and after correction for social desirability.

All three instruments that were used report a factorial structure that is more comprehensive than found in the interbattery factor analysis of the current study. What could be the reasons for the more simplified structure found here? An interbattery factor analysis identifies variance shared by the three instruments. The focus on shared variance implies that personality features that are measured in only one questionnaire remain uncovered in our analysis. Still, a content analysis of the items and scales of the three instruments would lead to the delineation of more common factors. For example, all three inventories have

scales that could merge into a shared Conscientiousness factor. Also, personality characteristics that are part of the OPP and 15FQ+ only, could have emerged as a separate factor. For example, both questionnaires contain subscales that measure flexibility. The main reason for not finding more common factors may be the low internal consistencies, particularly of the 15FQ+ (see Table 5.1). These low internal consistencies may also be the reason why we did not identify all factors of the Five-Factor Model. On the other hand, Cheung (2006) finds that the original facets of the Conscientiousness factor of the Five-Factor Model do not all load on the Conscientiousness factor; For example, Discipline loaded on Interpersonal Relatedness, whereas Dependability loaded both on Conscientiousness and Emotional Stability. Further research is required to examine whether Conscientiousness has a more narrow meaning in South Africa (which is demographically dominated by Black Bantu-speaking groups) than in western countries.

We also found that Extraversion split into two dimensions of dominance and sociability, which has not been seen in previous structural investigations of the Five-Factor Model. Although there is strong evidence for the link between dominance and sociability in many cultures (McCrae & Allik, 2002), the link may not be universal. Concepts like dominance and assertiveness are relatively underrepresented in the Chinese Personality Assessment Inventory, which was developed using an emic strategy. It may also be noted that in the Eysenck tradition, Extraversion only involves sociability and need for stimulation. In our view, a strong link between sociability and dominance is not very likely among Black South African groups. The concept of "Ubuntu" which focuses on communalism and interdependence ("I am because we are") plays an important role in these groups (Bewaji & Ramose, 2003). A core aspect of Ubuntu is that a person can only be a person through interaction with others. Important values associated with Ubuntu are group solidarity, conformity, compassion, respect, human dignity, and collective unity. Respect is reciprocal irrespective of race, ethnicity, class, age, and gender. Ubuntu requires one to respect others if one is to respect oneself. It is fair to assume that expressions of sociability and relatedness are under strict normative control in the Black groups. As a consequence, the relationship between sociability and dominance may be less salient in these groups.

We do not see our findings as conflicting with evidence about the universality of the Five-Factor Model and Giant Three. Personality can be studied at different levels. There is an important school of thought in cross-cultural psychology, called "universalism", which holds that at a high aggregation level

many psychological phenomena are identical across cultures; examples are the structure of personality and intelligence (Berry, Poortinga, Segall, & Dasen, 2002). However, specific manifestations of the underlying structure may not be universal, as cultures have non-overlapping norms with regard to the appropriateness of expressions of underlying traits, such as sociability. So, broad labels may be identical in these contexts but the way these are expressed may be different across cultures. In a similar vein, Marsella, Dubanoski, Humada, and Morse (2000) argue that "it may well be that there are a limited number of biologically or socially determined behavior dimensions but culture variations may shape their display patterns, situations in which they are elicited, the interpersonal responses to them, their utility or value in behavioral description, and the meaning they are assigned" (p.60).

A limitation of the current study is that we developed a new personality measure by making use of three English-language measures, two of which (the 15 FQ+ and OPP) were developed for use in western countries and adapted for South Africa. The third personality measure (the BTI) was developed in South Africa, but is based on an imported model, namely the Five-Factor Model. The question arises whether these measures uncover all personality factors that are important in the South African context (e.g., the concept of Ubuntu). We see the Chinese Personality Assessment Inventory as an important development in indigenous personality psychology. A similar development is required within the South African context to enhance our understanding and assessment of personality structure within the African context. The South African Personality Inventory Project (SAPI) has recently been undertaken (see, Meiring, 2006, for a full description of this project). The project aims at developing a single, unified personality inventory for South Africa that takes into consideration both universal and unique personality factors to be found across the various language groups in South Africa. It will be interesting to determine to what extent the factors identified in the Five-Factor Model, current study, the Chinese study, and the new indigenous study of South African personality show similarities and differences.

Author Note

We would like to acknowledge the feedback of Marise Ph. Born and Nicola Taylor on an earlier draft of the manuscript.

References

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of South Africa, Pretoria, South Africa.

Barrett, P. T., Petrides, K. V., Eysenck, S. B. G., & Eysenck, H. J. (1998). The Eysenck Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality and Individual Differences, 25,* 805-819.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology. Research and applications.* Cambridge, UK: Cambridge University Press.

Bewaji, J., & Ramose, M. (2003). The Bewaji, Van Binsbergen and Ramose debate on 'Ubuntu'. *South African Journal of Philosophy, 22,* 378-414.

Browne, M. W. (1979). The maximum likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology, 32,* 75-86.

Browne, M. W., & Tateneni, T. (1997). Noniterative estimation for the multiple battery factor analysis model. *Behaviormetrika, 24,* 3-18.

Chan, W. Ho, R. M., Leung, K. Cha, D. K-S., & Yung, Y-F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4,* 378-402.

Cheung, F. (2006). A combined emic–etic approach to cross-cultural personality test development: The case of the CPAI. In Q. Jing, H. Zhang, & K. Zhang (Eds.), *Psychological science around the world* (Vol. 2, pp. 91-103). London: Psychology Press.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross-Cultural Psychology, 27*,118-199.

Cheung, F. M., Leung, K., Zhang, J. X., Sun. H. F., Gan, Y. Q., Song, W. Z., & Xie, D. (2001). Indigenous Chinese personality constructs. *Journal of Cross-Cultural Psychology, 32*, 407-433.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69,* 979-1006.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69,* 979-1006.

Church, A. T., & Katigbak, M.S. (1988). The emic strategy in the identification and assessment of personality dimensions in a non-Western culture:

Rationale, steps, and a Philippine illustration. *Journal of Cross-Cultural Psychology, 19,*140-163.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Costa, P. T., & McCrea, R. R. (1992). *NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Cudeck, R. (1982). Methods for estimating between-battery factors. *Multivariate Behavioral Research, 17,* 47-68.

De Raad, B., & Peabody, D. (2005). Cross-cultural recurrent personality factors: Analysis of three factors. *European Journal of Personality.19,* 451-474.

De Raad, B., & Perugini, M. (2002). *Big five assessment.* Seattle, WA: Hogrefe and Huber.

De Bruin, G. P., Schepers, J. M., & Taylor, N. (2005, June*). Construct validity of the 16PF 5th Edition and the Basic Traits Inventory.* Paper presented at the 8th Annual Industrial Psychology Conference, Pretoria. South Africa.

De Bruin, G. P. (2005, August). *Cross-cultural psychological assessment in Africa: Lessons learnt in South Africa.* Paper presented at the 8th European Conference of Psychological Assessment, Budapest.

Eysenck, H. J. (1970). *The structure of human personality* (3rd ed.) London: Methuen.

Eysenck, H. J. (1992). Four ways five factors are not basic. *Personality and individual Differences, 13,* 667-673.

Government Gazette, Republic of South Africa, Vol. 400, no 19370. Cape Town, 19 October 1998.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessment: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11,* 147-157.

Harkness, J. A. (2003) Questionnaire translation. In J. A. Harkness, F.J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35-56). New York: Wiley.

Horn, B. S. (2000). *A Xhosa translation of the revised NEO Personality Inventory.* Unpublished master's dissertation, University of Port Elizabeth, Port Elizabeth, South Africa.

Heuchert, J. W. P., Parker, W. D., Strumf, H., & Myburg, C. P. H. (2000). The Five-Factor Model for African college students. *American Behavioral Scientist, 44,* 112-125.

Krug, S. E., & Johns, E. F. (1986). A large-scale cross-validation of second-order personality structure defined by the 16PF. *Psychological Reports, 59,* 683-693.

Lin, T. Y., & Church, A. T. (2004). Are indigenous Chinese personality dimensions culture-specific? An investigation of the Chinese Personality Assessment Inventory in Chinese American and European American samples. *Journal of Cross-Cultural Psychology, 35,* 586-605.

Marsella, A. J., Dubanoski, J., Humada, W. C., & Morse, H. (2000). The measurement of personality across cultures: Historical, conceptual, and methodological issues and considerations. *American Behavioural Scientist, 44,* 41-62.

McCrae, R. R., & Allik, J. (2002) (Eds.). *The five-factor model of personality across cultures.* New York: Kluwer Academic/Plenum Publishers.

McCrae, R. R., & Costa, P. T. Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509-516.

McCrae, R. R., Costa, P. T. Jr., Del-Pilar, G. H., Rolland, J. P., & Parker, W. D. (1998). Cross-Cultural Assessment of the Five-Factor model: The Revised NEO Personality Inventory. *Journal of Cross Cultural Psychology, 29,*171-188.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its application. *Journal of Personality, 60,*175-215.

Meiring, D. (2006). *SAPI Project.* Retrieved May 31, 2006 from the Web site http://www.meiringd.co.za/sapi project.htm.

Meiring, D., Van de Vijver, F. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality measures in South Africa. *South African Journal of Industrial Psychology, 31,* 1–8.

Meiring, D., Van de Vijver, F. J. R., & Rothmann, S. R. (2006). Bias in an adapted version of the 15FQ+ questionnaire in South Africa. *South African Journal for Psychology, 36, 340-356.*

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376-398.

Paunonen, S. V., & Ashton, M. C. (1998). The structured assessment of personality across cultures. *Journal of Cross-cultural Psychology, 29,* 150-170.

Paunonen, S. V., Zeidner, M., Enggvik, H., Oosterveld, P., & Maliphant, R. (2000). The nonverbal assessment of personality in five cultures. *Journal of Cross-Cultural Psychology, 31*, 220-239.

Piedmont, R. L., Bain, E., McCrea, R. R., & Costa, T. C. Jr. (2002). The application of Five-Factor Model in a Sub-Saharan culture: The NEO-PI-R in Shona. In R. R. McCrae & J. Allik, (Eds.), *The five-factor model of personality across cultures* (pp. 155-173). New York: Kluwer Academic / Plenum Publishers.

Psytech (2002a). *15FQ+ Fifteen Factor Questionnaire technical manual.* Retrieved December 2, 2005 from http:www.psytech.co.uk/downloads/manuals /15fq+% B man.pdf.

Psytech (2002b). *Occupational personality profile, the technical manual.* Retrieved December 2, 2005 from http:www.psytech.co.uk/downloads/ manuals/oppman.pdf.

Ramsay, L. J., Taylor, N., De Bruin, G. P., & Meiring, D. (2005, May). *The big five personality factors at work: A South African validation study.* Poster presentation at the International Symposium on Personality at Work 2005, Lueneburg, Germany.

Schepers, J. M. (2003, June).*Overcoming the effects of differential skewness of test items in scale construction.* Paper presented at Annual Conference of the Society for Industrial and Organisational Psychology of South Africa (SIOPSA).Pretoria, South Africa.

Spence, B. A. (1982). *A psychological investigation into the characteristics of black guidance teachers.* Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

Tact, H. (1999). *The cross-cultural validity and compatibility of the Sixteen Personality Factor Questionnaire.* Unpublished master's dissertation, University of Pretoria, Pretoria, South Africa.

Taylor, N., & De Bruin, G. P. (2005). *Basic Traits Inventory: Technical manual.* Johannesburg: Jopie van Rooyen & Partners SA.

Taylor, I. A. (2000). *The construct comparability of the NEO PI-R Questionnaire for Black and White employees.* Unpublished doctoral thesis, University of the Orange Free State, Bloemfontein, South Africa.

Taylor, T. R., & Boeyens, J. C. A. (1991). *A comparison of black and white responses to the South African Personality Questionnaire.* Pretoria: Human Sciences Research Council.

Tredoux, N. (2004). *15FQ+ SA Reliability.* Retrieved March 1, 2004 from the Web site http://www.psytech.co.za/15FQPlus_reliabilities.htm.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika, 23,* 111-136.

Tyler, G. (2002). *A review of the 15FQ+ Personality Questionnaire.* Pulloxhill, UK: Psychometrics Limited.

Van de Vijver, F. J. R. (2003). Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143-155). New York: Wiley.

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1, 89-99.*

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33,141-156.*

Van Hemert, D. A., Van de Vijver, F. R. J., Poortinga, Y. H., & Georgas, J. (2002). Structure and score levels of the Eysenck Personality Questionnaire across individuals and countries. *Personality and Individual Differences, 33, 1229-1249.*

Chapter Six
General Discussion and Conclusion: Integrating the Main Findings

President Nelson Mandela told an excited, cheering crowd of 60,000 people attending his inauguration at the Union Buildings in Pretoria in 1994 that the "rainbow nation" had at last achieved its political emancipation and that "never, never, and never again shall it be that this beautiful land will experience the oppression of one by another". He concluded:

> *"I stand firm in the belief that we are one country, one nation, whether we are Coloureds, Indians, White or Africans – that is what we must promote in this country from now on."* (*Business Day*, 1994).

Psychological assessment can be seen a product of the history and the dynamics of the South African society where Apartheid had played a major role in influence testing practices. Democratization brought along new challenges for psychologists in establishing a new paradigm for assessment practices. New legislation in the form of the Employment Equity Act No. 55 of 1998 (Government Gazette, 1998) has also set out new rules whereby psychologists using psychological measures have to indicate that they adhere to the regulations of the legislation. The way psychologists respond to the legislative challenges will largely shape the future of psychological assessment in South Africa.

The current project attempted to add to the body of knowledge in the field of cross-cultural assessment in South Africa and at the same time to give insights to psychologists as to how to rise to these new challenges. The study made use of large samples representing respondents of all the cultural groups in South Africa. I was interested in the question of whether the test battery used by the South African Police Service complies with the Employment Equity Act, according to which psychological tests can only be used if they are unbiased.

The research project included four empirical studies. The aim of the first two studies was to investigate the structural equivalence and the influence of test adaptations on equivalence (e.g., internal bias). The third study's main aim was to address external bias in examinations of two independent cohorts. In the fourth study I wanted to go a step further by examining to what extent it is possible to compile an appropriate instrument by combining existing measures. The four

studies together provide a test of how adequate existing instruments are with regard to bias and equivalence within the South African context.

Main Findings

In the first study (Chapter 2) we examined the internal bias (construct, item, and method) of psychometric tests included in the selection process of applicants for entry level positions in the South African Police Service (SAPS). We were not able to detect any sizable bias in the cognitive instruments; the bias that was detected was negligible from a practical perspective. The findings for the personality measures (e.g., 15FQ+) were quite different; various scales revealed construct bias, but item bias was small. We did not find any method bias. A major finding of the study was that several personality scales revealed low internal consistencies, notably in the Black groups. We concluded that in high-stake testing situations the 15FQ+ was not an adequate measure for the South African context.

The first study suggested various ways to adapt the 15FQ+ in order to deal with the serious problems associated with the construct bias and low internal consistency of the original 15FQ+ instrument. We assumed that a major reason for these problems was the relatively complex language of the instrument for non-native speakers. In the second study (chapter 3) we adapted the 15FQ+, by changing various words and sentence constructions. The adapted version of the 15FQ+ produced less construct and item bias than the original version. The main finding of this study was that adaptations of the instrument essentially solved the cross-cultural equivalence problems, but low internal consistencies prevailed especially in the large Black group. Despite our efforts to adapt the instrument, the measure still remains unsuitable for use in the police test battery due to the low internal consistencies.

In study 3 (chapter 4) we investigated the external bias of the police test battery and focused on the relationship between two observed variables (e.g., cognitive and personality measure) and a criterion (e.g., police training). We evaluated the test battery for compliance with the technical standards for predictive bias as put forward in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures, developed in the U.S.A. and also adopted in South Africa. Within the study we examined two independent cohorts to test whether the police battery would show any predictive bias. In the first cohort we found that both the predictor measures and training criteria were reliable and equivalent. Good fit indexes were found for a regression model with an equal slope and intercept for each of the different groups. Range restriction did

not have any sizable influence on the regression results. In the second cohort we subsequently tried to improve on the results of the first study (e.g., predictive power and differential bias) by employing adapted personality measures. The adapted versions of the personality questionnaires yielded further improvements; we found an overall increase in the multiple correlation of the second cohort. Our study concluded by indicating that neither cohorts showed evidence of internal or external bias. The test battery further showed compliance with technical standards put forward in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures. Based on the results of these two cohorts, we concluded that the SAPS test battery complied with the requirements of the Employment Equity Act in South Africa.

In the last study (Chapter 5) we investigated the commonalities of three comprehensive measures of personality traits to develop a new instrument that complies with legislative requirements. Commonalities of three separate comprehensive personality measures (e.g., 15FQ+, OPP and BTI) were investigated by performing an interbattery factor analysis. The results of the study reported a new personality instrument with four stable factors with good psychometric properties. Structural equivalence of the four factors was found across all the cultural groups. Although many items of the new instrument showed some bias, the impact of the bias on the size of cross-cultural differences was negligible from a practical perspective.

Conclusion

Our findings provide a broad picture of the interplay of both internal and external bias of psychological measures in a multilingual and multicultural society such as South Africa. The current research project is the first comprehensive study undertaken in South Africa that explored both internal and external bias. Although the outcomes of our project were very favorable, especially Study 3 (e.g., predictive bias), indicating that the SAPS selection procedure meets the legal South African standards, we need to carefully analyze whether the selection procedure can be said to be fair in every possible sense. In the literature it appears that there is considerable confusion about the use and implications of the terms of test bias and test fairness (Koch, 2005).

Fairness, which is also a requirement of Section 8 of the Employment Equity Act, refers to a social rather than a psychometric concept. Fairness (or the lack thereof) is not the result of the assessment instrument or predictor, nor is it a property of the assessment procedure used (SIOPSA, 2005). Fairness is a consequence of the influence of all the variables that play a role in the final

decision based on an assessment procedure. Fairness is the total of all the variables that play a role or influence the final personnel decision. This may include the test, predictor, and integration of data, recommendations based on these data or the final decision made by line management. It must thus be regarded as an important issue, especially in the case of large-scale high-stake testing such as in the SAPS.

Bias refers to "any construct-irrelevant source of variance that results in systematically higher of lower scores for identifiable groups" (SIOPSA, 2005, p. 26). Koch (2005) points out that bias refers mainly to technical issues associated with construct as well as predictive validity. While an unbiased test will not necessarily be seen as fair, fairness issues are not directly dealt with during the assessment of bias. Therefore, bias does not guarantee fairness in a cultural context in which score differences on predictors are affected by confounding intergroup differences such as access to good education or people's understanding of English which is either their second or third language.

A recurring theme in this project was the use of English as medium of assessment across multilingual groups and its influence on the psychometric properties of tests. English can be used in a dual way in the assessment procedure; it can be the target of testing (as was the case in our cognitive tests, e.g. the Reading Comprehension Test), but it can also be the vehicle of the testing (as was the case in our personality questionnaires) (Yeld, 2001). The question is whether it is fair to use English as a medium of assessment in a multilingual society such as South Africa where we have a majority group that are less fluent in English. Cele (2004), for instance, indicates that only 9% of the South African population speaks English as a first language, and about 14% Afrikaans and that only 20% of the population is fluent in English. The use of English is usually more problematic in questions of internal bias than of external bias, because actual job performance requires a high level of English proficiency. In that case the bias is shared by the predictor (test battery) and the criterion. However, if the test battery is used in an internal bias examination, level of English can easily become a confounding variable, as incorrect inferences about constructs measured by the test battery, such as personality traits, are easily made when applicants do not fully grasp the meaning of the items. In this project we have reported that language was the main issue that had an impact on test bias.

Language usage in South African is closely linked with historical developments of the country and the language policies in the education system that was adopted by the government of the day. The obvious conclusion that one

comes to is that the coercive and authoritarian imposition of language policies, both by British and by Afrikaner nationalists not only failed, but left behind a deeply divided society. Language policies in education are an emotive issue because they have been used to preserve the cultural and linguistic identities of some groups and not of others; it has been used to segregate and discriminate against various race groups. In the previous political dispensation, the South African Apartheid government did not only impose a bilingual (Afrikaans/English) education policy, it also tried to impose mother tongue instruction policy on African schools. Resentment of these policies triggered the 1976 Soweto uprising which spread to other parts of the country, eventually reaching the proportions of a civil war. Significantly, when the government gave in and agreed to one medium, to be decided by the school from Standard Five onwards, the overwhelming choice was for English rather than Afrikaans. Language was the basis on which people were classified per ethnic group. For African people, this meant that their various mother tongues were crucial in classifying which group they belonged to.

The use of English in Black schools is grosssly inadequate. In their two years of schooling, African pupils study through the medium of vernacular and do English only as a subject. In the next two years (grades three and four), English is introduced as a subject and from the fifth grade onward, besides taking English as a subject, pupils use it as medium of instruction (Gough & England, 1991). The policy is based on the principle that the child's first language facilitates a smooth transition between the informal learning of the home or preschool and the formal education of junior primary school. Research co-ordinated by Macdonald (1990) for the Human Sciences Research Council (HSRC), is probably the most significant piece of research conducted on the role of language in education in South Africa. A central finding of Macdonald's research was that school learners who change the language of instruction before they have sufficiently developed and learnt the new target language of learning would not succeed. In a recent study focussing on theory and practice of language and education models in Africa, Heugh (2006) drew the following conclusions:

- Early education in Africa languages is a good thing, but if its education benefits are to be of a lasting value, then mother-tongue and mother-tongue medium education need to continue at least to the end of the sixth year and preferably longer.
- Mother tongue literacy and oral language development need to be developed to the level that written text and oral language used

for learning and teaching mathematics, science, history, and geography can be understood and actively used by the learner.

- If a switch in medium of instruction takes place before learners have developed high levels of written as well as spoken proficiency, then the learning process across the curriculum will be interrupted and learners will fall behind their peers.
- Under optimal conditions, it takes 6-8 years to learn a second language sufficiently well to use it as a medium of instruction.

Educators and pupils from disadvantaged communities work against great odds in South Africa despite all the efforts of the new government. This is largely because many educators are, besides being ill-trained, not proficient to use English as a language of instruction. Shortage of classrooms, overcrowding, and long distances that pupils need to walk to schools also contribute to the problem. Kilfoil and Van der Walt (1997) indicated that rural pupils have more difficulty with English as it is not spoken much in their everyday environment. The authors state that since pupils in rural schools have few opportunities to speak English outside the classroom, the likelihood of their learning English successfully is rendered all the more difficult.

From the above, a clear picture emerges that English as a medium of test administration will be problematic in South Africa. We can only come to the conclusion that testing in English seems to be unfair. The development of culturally appropriate measures in the multicultural, multilingual South African society is thus fraught with difficulties. According to Shuttleworth-Jordaan (1996), various cultural groups in South Africa are at various stages of westernization, and given the linguistic and educational diversity of its peoples, the development of culturally relevant tests will be extremely difficult and almost an unattainable goal. South Africa is not simply a multicultural society, it is a multicultural society in which acculturation of many kinds is taking place and in which a new citizenship is actively encouraged.

The question of imported, adapted, and locally developed tests is an important question in this project. I came to the conclusion that importing tests (e.g., personality tests) is problematic if you use them in their original form. Research by Abrahams (1996, 2002) also confirmed this. Adapting imported measures for the South African context resulted in small gains, although chapter 3 of this project shows that the limits of what can be reached were quickly approached. Developing a new instrument in South Africa seems to be a more viable option. An example of this is the work by Taylor and De Bruin (2005) who

developed the Basic Traits Inventory (BTI). The BTI was specifically designed for cross-cultural use in South Africa with careful consideration of item content, structure, and presentation. The authors used a top-down approach using the Five-Factor Model as their structure of departure. This instrument worked well within the cross-cultural context and showed good psychometric properties. Another example, namely the South African Personality Inventory (SAPI) project, was introduced in Chapter 5 as a new approach. The proposed project aims at developing a single, unified personality inventory for South Africa that takes into consideration both universal and unique personality factors to be found across the various cultural groups in South Africa. This project uses a bottom-up approach and does not start from a well-known conceptualization of personality, but rather tries to start from an everyday conceptualization of personality as found in South African groups.

Practical Implications

The field of psychological test use, development, and adaptation in South Africa faces many challenges at present. Foremost among the challenges is that culturally appropriate tests, which meet stringent psychometric standards, are needed for all age groups in our multicultural society if psychological assessment practitioners are to succeed in employing fair testing practices. Linked to this is the challenge of having various language versions of tests so that test-takers in the multilingual South African society can be assessed in the language in which they are most proficient. Intensive, large-scale test development, adaptation and revision projects need to be urgently undertaken if South African psychological assessment practitioners are to rise to the challenge of performing ethically and culturally sound assessment (Foxcroft, Paterson, Le Roux, & Herbst, 2004).

The Human Sciences Research Council (HSRC), which was the largest developer and supplier of psychological tests until the early 1990s, has been restructured and it is still uncertain as to what role, if any, it could and should play in test development in the 21st century. Since the demise of the HSRC as a major test developer, smaller companies have sprung up to import, develop and supply tests, but there is no body or organization to coordinate test development activities.

The key driver for a new research agenda in terms of psychological assessment is clearly the new constitution and stronger demands for cultural appropriateness of psychological tests culminated in the promulgation of the Employment Equity Act 55 of 1998, Section 8. There is an urgent need for measuring instruments that meet the Employment Equity Act requirements and

can be used for all the cultural and language groups in South Africa. It is quite clear that there is a need for a single organization to take ownership of psychological assessment in South Africa, particularly relating to issues of quality, good testing practices, culturally appropriate tests and psychological assessment functioning within a framework of an internationally agreed set of guidelines. I am of the opinion that South Africa is in need of a professional body that can take up this role. Although the Professional Board for Psychology in South Africa has established a Psychometrics Committee to oversee psychological testing, it has failed to establish a national test development, adaptation and revision agenda that could drive and coordinate test development efforts undertaken by the various smaller companies and research teams at universities. A good example of such an organization is the British Psychological Society (BPS), Psychological Test Centre (PTC). In January 2003, the PTC was formed to consolidate activities in psychological testing to include areas of educational and clinical testing and the publishing and dissemination of advisory statements in addition to guidelines and other literature on standards for the construction, use, and availability of tests; the training requirements and qualifications of test users are considered in the United Kingdom (UK). An agency like the PTC is needed in South Africa to reach the following aims and objectives (Psychological Testing, 2006):

- To provide a quality professional service of support and advice on matters of psychological testing to members of the profession and of the public
- To use the processes of user certification and test reviewing to support good practice and the development of quality tests
- To provide a psychological testing centre on-line to all the key sources of information relating to tests and testing.

Objectives

- To implement steering committee on test standards policy and disseminate standards on best practice in testing and test use
- To establish and maintain a comprehensive psychological testing website and the psychological testing centre on-line
- To manage all test user certification procedures
- To manage the verification of test user competence assessments.

In my view, this agency will have to deal with the following issues:

- It seems that psychological assessment in South Africa is still haunted by its legacies of the past. If we do not do something about it we will still haunt us in the future. Assessment practitioners must not abdicate their assessment function. In the early 1970s and 1980s this happened in Europe, with little training taking place, no psychometric research undertaken and few measures developed. The result was that other professionals took over our role and it was difficult to regain control over who is permitted to use psychological measures. Psychologists in South Africa cannot afford it.
- The question can also be raised as to whether psychology as a profession in South Africa is ready for the challenge implied by the Employment Equity Act. It is probably fair to say that the law is ahead of the daily practice here and that, to date no single country can live up to the expectations and demands raised by the Act. One of the main goals of the assessment profession in South Africa is to bring current practices in line with legal demands, for example by developing new instruments and validating existing instruments for use in multicultural groups. On the short term the Act may be seen as a threat to psychologists; on the longer term the Act may enhance the professional level of psychological practice by putting multicultural assessment on the agenda of the profession and by stimulating the development of new tests and even new testing practices.
- The development of psychological tests is highly specialized and should be undertaken by teams of experienced measurement experts. Until the 1990s the Human Sciences Research Council (HSRC) almost exclusively developed or adapted the psychological tests used in South Africa. However, while there was considerable test development expertise at the HSRC, little emphasis was placed on training test developers in postgraduate psychology programs, which meant that test development capacity was not built among younger researchers. The situation was further compounded when, during the process of transforming and restructuring itself in the mid 1990s, the development of psychological tests was not the focus of the HSRC anymore. Many of the experienced test developers retired, were redeployed to other positions in the organization, took up positions at academic institutions, or emigrated. However, it is unfortunate that at this critical moment when psychological test development stands at the threshold of a new era in which new tests

should be developed from a multicultural rather than a monocultural perspective, there is a critical shortage of research and competence in this area in South Africa.

According to Foxcroft et al. (2004), few research data are available related to test use patterns and the needs of assessment practitioners which could aid in the establishment of a national agenda. In the absence of a national agenda a major project was undertaken by the HSRC in 2004 to conduct a need analysis on psychological assessment in South Africa among practitioners. A tentative agenda was generated out of the need analysis to guide the continued development of psychological testing in South Africa. The key aspects of the agenda from the report of Foxcroft et al. (2004) were the following:

- Develop a clear description of a psychological test and how it is differentiated from other types of tests.
- Disseminate user-friendly information on what a psychological test is and who may use such a test as an awareness raising exercise among those stakeholders external to the psychology professions who use psychological test information.
- Clearly delineate the purposes of test use and the scope of practice for professionals in the various registration categories and re-align training programs accordingly.
- Urgently review the requirement that practitioners should only use tests registered with the Professional Board for Psychology and shift the emphasis to requiring that practitioners use high quality, culturally appropriate tests.
- Review the test classification process and find mechanisms to complete the process with greater speed.
- Introduce a comprehensive test review system using a standardized format and provide practitioners with easy access to the review information.
- Provide appropriate continuing professional development activities related to advances in the field, psychological test use and assessment practice.
- Existing national and international tests identified in the present project need to be urgently adapted, revised or updated.
- New culturally and linguistically appropriate tests need to be developed to fill gaps in the toolkits of assessment practitioners.

- Develop competency standards for all levels of assessment practitioners and revise training programs where necessary so that they can produce practitioners with the desired competencies.
- Develop training programs to build test development expertise; Black test developers in particular are needed.
- Develop a Code of Practice for test developers, publishers, and distributors.
- Establish the role that professional societies need to wish to play.
- Develop a model with respect to who will be responsible for developing tests and who might coordinate test development.

A comprehensive survey of the test-use patterns and needs of psychological assessment practitioners has never been undertaken in South Africa the project by Foxcroft et al. (2004) is one of the first constructive efforts to put psychological testing on the national agenda in South Africa.

In conclusion, the way in which psychologists respond to new challenges will largely shape the future destiny of psychological testing in South Africa. The current challenge faced by testing in South Africa has the potential to stimulate the developers and users of psychological tests to new heights in their quest to develop appropriate measures that can be used in fair and unbiased ways to the benefit of individuals.

References

Abrahams, F. (1996). *The cross-cultural comparability of the Sixteen Personality Factor Inventory (16PF).* Unpublished doctoral thesis, University of Pretoria, Pretoria, South Africa.

Abrahams, F. (2002). Fair usage of the 16PF (SA 92) in South Africa: A response to C. H. Prinsloo & I. Ebersohn. *South African Journal of Psychology, 32,* 58-61.

British Psychological Society. (2006). Psychological test centre/About us. Retrieved August 1, 2006, from http://www.psychtesting.org.uk/viewer.asp

Camilli, G. (1993). The case against item bias techniques based on internal criteria: do item bias procedures obscure test fairness issues? In P.W. Holland & H. Wainer (Eds.), *Differential item functioning (*pp. 397-413). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cele, N. (2004). Equity of access and equity of outcomes challenged by language policy, politics and practice in South African higher education: The myth of language equality in education. *South African Journal of Higher Education, 18(1),* 38-55.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-220). London: Collier Macmillan publishers.

Foxcroft, C. D., Paterson, H. Le Roux, N., & Herbst, D. (2004). *Psychological assessment in South Africa: A needs analysis.* Pretoria: Human Sciences Research Council.

Gough, D., & England, V. (1991). *Medium of instruction in South Africa: Historical overview.* Unpublished report. Johannesburg: NEPI.

Government Gazette, Republic of South Africa, Vol. 400, no. 19370. Cape Town, 19 October 1998.

Heugh, K. (2006). Theory and practice – Language education models in Africa: Research, design, decision-making, and outcomes. In H. Alidou, A. Boly, B. Brock-Utne, Y. S. Diallo, K. Heugh, & H. E. Wolff (Eds.), *Optimizing learning and education in Africa: The language factor. A stock-taking research on mother tongue and bilingual education in sub-saharan Africa* (pp. 56-84). Paris: Association for the Development of Education in Africa (ADEA).

Kilfoil. W. R., & Van der Walt. C. (1997). Learn to teach: English language teaching in a multilingual context (3rd ed.). Pretoria: Van Schaik.

Let's build a great SA, says Mandela. (1994, May 11). Business Day, pp.1.

Koch, E. (2005). Evaluating the equivalence, across language groups, of a reading comprehension test used for admission purposes. Unpublished doctoral thesis, University of Port Elizabeth, Port Elizabeth, South Africa.

Macdonald, C. A. (1990) *Crossing the threshold into standard three in Black education: The consolidated main project of the Threshold Project.* Pretoria: Human Sciences Research Council.

Shuttelworth-Jordaan, A. B. (1996). On not reinventing the wheel: A clinical perspective on culturally relevant test usage. *South African Journal of Psychology, 26,* 184–192.

Society for Industrial and Organizational Psychology in South Africa (2005). *Guidelines for the validation and use of assessment procedures for the workplace* (3rd ed.). Pretoria: Author.

Taylor, N., & De Bruin, G. P. (2005). Basic Traits Inventory: Technical manual. Johannesburg: Jopie van Rooyen & Partners SA.

Yeld, N. (2001). Assessment, equity and language of learning: Key issues for higher education selection in South Africa. Unpublished doctoral thesis, University.

Summary

If the history of South Africa is in large part one of increasing racial divisiveness, today it can be seen as the story of - eventually - a journey through massive obstacles dealing with tremendous diversity towards the creation, of a single nation; the dream of unity and common purpose can now be realised. Legislation in the form of the Employment Equity Act, which stipulates that no psychological test that is biased against any cultural group can be used in South Africa, has transformed the landscape of the use of psychological measurement in South Africa. The new legislation poses challenges to psychologists who make use of these devices in their practices or organisations. The current study took place in the South African Police Services which is a public service entity that employs psychologists who make use of these instruments to select police applicants on a yearly basis.

The current project attempted to investigate to what extent current assessment practices can live up to the legal requirements, using models of bias and equivalence that have been proposed in cross-cultural psychology. The main question to be answered in this study was to what extent the South African Police Services selection battery complies with the Employment Equity Act. This thesis attempted to answer this question by examining the bias and equivalence of the entry level test battery of SAPS. A test of bias in this selection battery can be seen as a good test of compliance because the applicants come from all major South African cultural groups and the test battery uses types of instruments that are commonly used in South Africa for selection purposes.

In the *first chapter,* the societal context in which psychological assessment is embedded was described. It was followed up with a section on cross-cultural assessment within a multicultural society such as South Africa. A theoretical model for the different components of bias and equivalence were described and discussed. A model is described in which bias (i.e., systematic distortions in test scores due to cultural or linguistic factors) can derive from the construct of an instrument (construct bias), the method to assess the construct (method bias), or to specific items (item bias). The literature showed that there are almost no empirical studies of bias in South Africa.

The *second chapter* describes the first study, which was aimed at assessing construct, item, and method bias of cognitive and personality tests in South Africa. The study had two major research aims, the first aim was to examine bias at the level of constructs and items bias in two cognitive tests and a personality test that were administered to select entry-level police officials for the

SAPS. The second aim was to examine method bias studying the influence of cognition and social desirability on the 15FQ+. With regard to the first aim, it was found that the cognitive instruments produced very good construct equivalence and low item bias. However, various scales of the personality questionnaire revealed construct bias in several ethnic groups. The item bias in the personality scales was small. With regard to the second research aim, it was found that method bias did not have any impact on the (small) size of the cross-cultural differences in the personality scales. It was concluded that the 15FQ+ was not suitable as an instrument in the South African multicultural context because of the low internal consistencies of some scales and the lack of construct equivalence. The authors concluded that when applying personality measure from abroad in the South African context we need to be sensitive to the huge socioeconomic and development differences that do exist.

The *third chapter* reports a study that was a followup study to chapter two study. The study in the second chapter suggested various ways to adapt the 15FQ+ so as to increase its cross-cultural suitability. The study of the third chapter examined bias in an adapted version of the 15FQ+ that was part of a test battery used to recruit new police officials for the South African Police Services (SAPS) in 2001. The study reported that the adapted version of the 15FQ+ produced less construct and item bias than the original version. The main finding was that the adaptations of the instrument essentially solved the cross-cultural equivalence problems, but low internal consistencies prevailed especially in the large Black group. Despite the researchers' efforts to adapt the instrument, the measure still remains unsuitable for use in the police test battery due to the low internal consistencies.

The *fourth chapter* reports a study that focused on the external bias, examining the relationship between observed variables (e.g., test battery) and a criterion (training results). Two cohorts of trainees were examined, investigating whether the police battery showed any predictive bias and whether the test battery complied with technical standards for predictive bias as proposed in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures in the USA. The results of the first substudy showed that the predictive validity of the police battery was encouraging, subsequently in the second substudy the researchers tried to improve on the results (e.g., predictive power and differential bias) by employing adapted personality measures. The main results of the second study showed that adapted versions of the personality questionnaires yielded further improvements. The main conclusion of the two studies was that neither cohort showed evidence of internal or external bias. The

test batteries showed compliance with technical standards put forward in the fourth edition of the Principles for the Validation and use of Personnel Selection Procedures. It was concluded that the SAPS test battery adhered to legislative requirements of the Employment Equity Act in South Africa.

The *fifth chapter* investigated the commonalities of three comprehensive measures of personality traits to develop a new instrument that comply with legislative requirements. The three measures were combined into a single new personality instrument. The new instrument, which had good psychometric properties, yielded four stable factors. Structural equivalence of the four factors was found across all the cultural groups. Although many items of the new instrument showed some bias, the cross-cultural differences were so small to be negligible from a practical perspective. The newly developed instrument complied with the legislative requirements.

The *final chapter* summarized the main findings, conclusions and practical implications of the study. Although the SAPS selection battery complied with the legislative requirements of the Employment Equity Act, the question of whether the battery can be applied fairly in multicultural society is not easy to answer. English as a medium of test administration was problematic and testing in this medium in South Africa seems to be unfair. Developing cultural appropriate test and different language versions seems to the main challenges ahead of psychologists. A single testing agency for South Africa has been proposed as a catalyst from where a national agenda on psychometric testing can be coordinated and guided. It was concluded that the way in which psychologists respond to the new legal and societal challenges will largely shape the future and destiny of psychological testing in South Africa.

Nederlandse Samenvatting (Dutch Summary)

Als de historie van Zuid-Afrika er in belangrijke mate een is van toenemende raciale verdeeldheid, kan het verhaal van vandaag de dag –eindelijk— gezien worden als een reis via enorme obstakels om met de enorme diversiteit om te gaan naar één land; de droom van eenheid en een gemeenschappelijk doel kan nu gerealiseerd worden. Wetgeving in de vorm van de Employment Equity Act, die bepaalt dat geen enkele psychologische test die een vertekening tegen enige culturele groep vertoont gebruikt mag worden in Zuid-Afrika heeft het testlandschap in Zuid-Afrika veranderd. De nieuwe wetgeving is een uitdaging voor psychologen die gebruik maken van psychologische instrumenten in hun praktijk of organisaties. De studie die hier beschreven wordt vond plaats binnen de South African Police Services. Dit is een organisatie die diensten aanbiedt. De organisatie heeft psychologen in dienst die met behulp van deze instrumenten jaarlijks sollicitanten selecteren voor een baan bij de politie.

Dit project probeerde na te gaan in hoeverre de huidige praktijk van psychologische tests en testgebruik in overeenstemming is met de wettelijke voorschriften; deze vraag is onderzocht met behulp van modellen voor vertekening en equivalentie die in de crossculturele psychologie ontwikkeld zijn. De belangrijkste vraag van de studie was in hoeverre de testbatterij van de Zuid-Afrikaanse politie in overeenstemming is met de Employment Equity Act. Dit proefschrift probeerde een antwoord te geven op deze vraag door vertekening en equivalentie te bestuderen van de selectiebatterij van de Zuid-Afrikaanse politie. Een studie van de vertekening in deze batterij kan gezien worden als een goede toetsing van de vraag in hoeverre de procedure voldoet aan de wet omdat de sollicitanten uit alle belangrijke Zuid-Afrikaanse culturele groepen komen en de testbatterij soorten instrumenten gebruikt worden die veel gebruikt worden voor diagnostische doeleinden in Zuid-Afrika.

In het *eerste hoofdstuk* wordt de maatschappelijke context van psychologische diagnostiek beschreven, gevolgd door een sectie over crossculturele diagnostiek in een multiculturele samenleving als Zuid-Afrika. Er wordt een theoretisch model voor verschillende componenten van vertekening en equivalentie besproken. In het model kan vertekening (systematische verschuivingen in testscores als gevolg van culturele of linguïstische factoren) een gevolg kan zijn van het construct (constructvertekening), de methode om het

construct te meten (methodevertekening) of specifieke items (itemvertekening). In de literatuur zijn bijna geen empirische studies hierover in Zuid-Afrika te vinden.

Het *tweede hoofdstuk* beschrijft de eerste studie, die construct-, methode- en itemvertekening probeerde op te sporen in cognitieve tests en persoonlijkheidsvragenlijsten in Zuid-Afrika. De studie had twee doelen. De eerste was om construct- en itemvertekening te onderzoeken in twee cognitieve tests en een persoonlijkheidsvragenlijst die voorgelegd waren aan sollicitanten van de South African Police Services. Het tweede doel was om methodevertekening te bestuderen door de invloed van cognitie en sociale wenselijkheid op de 15FQ+ na te gaan. Met betrekking tot het eerste doel werd gevonden dat de cognitieve instrumenten vrijwel geen constructvertekening vertoonden en een goede betrouwbaarheid hadden. Verscheidene schalen van de persoonlijkheidsvragenlijst vertoonden echter wel constructvertekening in verscheidene culturele groepen. De itemvertekening in de persoonlijkheidslijsten was klein. Met betrekking tot het tweede doel werd gevonden dat methodevertekening geen invloed had op de toch al kleine crossculturele verschillen op de persoonlijkheidsvragenlijst. De conclusie was dat de 15FQ+ geen goed instrument was voor de Zuid-Afrikaanse context vanwege de lage betrouwbaarheden van enkele schalen en het gebrek aan constructequivalentie. De auteurs kwamen tot de slotsom dat we bij het gebruik van persoonlijkheidsinstrumenten van elders in de Zuid-Afrikaanse context goed moeten letten op de enorme verschillen in socio-economische status en ontwikkeling van culturele groepen.

Het derde hoofdstuk beschrijft een studie die een follow-up was van het vorige hoofdstuk. De studie in het tweede hoofdstuk bevatte allerlei aanwijzingen over mogelijke adaptaties van de 15FQ+ om de crossculturele bruikbaarheid van het instrument te vergroten. De studie die in het derde hoofdstuk beschreven wordt onderzocht vertekening in een geadapteerde versie van de 15FQ+ die in 2001onderdeel was van een testbatterij die gebruikt wordt om aspirant-politieagenten voor de South African Police Services (SAPS) te selecteren. De resultaten gaven aan dat de geadapteerde versie van de 15FQ+ minder construct- en itemvertekening vertoonde dan de oorspronkelijke versie. De belangrijkste bevinding was dat de adaptaties van het instrument de problemen met de crossculturele equivalentie nagenoeg oplosten, maar dat lage betrouwbaarheden een probleem bleven, vooral de in de zwarte groep. Ondanks de inspanningen van de onderzoekers om het instrument aan te passen is de lijst nog steeds niet bruikbaar om als selectie-instrument in de testbatterij van de politie ingezet te worden vanwege de lage betrouwbaarheden.

Het vierde hoofdstuk beschrijft een studie die vooral ingaat op externe vetrekening; de relatie wordt onderzocht tussen geobserveerde variabelen (zoals een testbatterij) en ene criterium (testscores na ene training). Twee trainingscohorten werden onderzocht; er werd nagegaan of de politiebatterij enige predictieve vertekening vertoonde en of de testbatterij voldeed aan de technische richtlijnen voor predictieve vertekening, zoals deze voorgesteld zijn in de vierde uitgave van de Principles for the Validation and Use of Personnel Selection Procedures in de Verenigde Staten. De resultaten van de eerste deelstudie gaven aan dat er weinig predictieve vertekening in de politiebatterij zat; vervolgens is in een tweede deelstudie met geadapteerde instrumenten gewerkt. Er werd nagegaan of de nieuwe instrumenten meer voorspellende waarde en minder vertekening zouden vertonen. De belangrijkste bevinding van de tweede studie was dat de nieuwe instrumenten inderdaad beter waren. De belangrijkste conclusie van de studie was dat de batterij noch interne noch externe vertekening vertoonde. De testbatterij voldeed aan de technische richtlijnen zoals die voorgesteld zijn in de vierde uitgave van de Principles for the Validation and use of Personnel Selection Procedures. De belangrijkste uitkomst van het onderzoek was dat de SAPS testbatterij in overeenstemming was met de wettelijke voorschriften van de Employment Equity Act in Zuid-Afrika.

Het *vijfde hoofdstuk* onderzocht de gemeenschappelijke elementen in drie persoonlijkheidsvragenlijsten die in overeenstemming zijn met de wettelijke voorschriften. De drie lijsten werden in één nieuw persoonlijkheidsinstrument gecombineerd. Vier factoren werden gevonden in het nieuwe instrument, dat goede psychometrische eigenschappen bleek te hebben. Structurele equivalentie van de vier factoren werd gevonden in alle culturele groepen. Hoewel veel items een beetje vertekening vertoonden, waren de crossculturele verschillen praktisch verwaarloosbaar. Het nieuw ontwikkelde instrument voldeed aan de wettelijke voorschriften.

Het laatste hoofdstuk bevat een weergave van de belangrijkste bevindingen, conclusies en praktische implicaties van de studie. Hoewel de selectiebatterij van de SAPS in overeenstemming is met de wettelijke vereisten zoals neergelegd in de Employment Equity Act, is de vraag of de batterij op eerlijke wijze toegepast kan worden in een multiculturele samenleving niet gemakkelijk te beantwoorden. Het gebruik van Engels bij het testen was problematisch en het gebruik ervan voor diagnostiek in Zuid-Afrika kan als oneerlijk aangemerkt worden. Het ontwikkelen van adequate tests in verschillende talen is een van belangrijkste taken van psychologen die voor ons liggen. Er wordt voorgesteld om één testorganisatie op te richten die een

nationale agenda op het terrein van psychometrisch testen kan coördineren en leiden. Er werd geconcludeerd dat de manier waarop psychologen reageren op de nieuwe wettelijke en maatschappelijke uitdagingen voor een belangrijk deel vorm zullen geven aan de toekomst en de bestemming van psychologisch testen in Zuid-Afrika.

Deon Meiring was born in 1967 in Carletonville, South Africa where he completed his elementary and high school education. He obtained both his B.A and B.A. Honours degrees in Industrial Psychology from the North-West University (Potchefstroom Campus). He completed his MA degree in Industrial Psychology in 1995 at the University of South Africa (UNISA). For the past 15 years he has been working as a Psychologist in the South African Police Services, Psychological Services. He is a member of the WorkWell, Research Unit for People, Policy and Performance of the North-West University, Potchefstroom Campus where he is responsible for the focus area "Personality in the Workplace" of which the SAPI project is one of his responsibilities. In 2002, he began his Ph.D. study at Tilburg University. He was supervised by Prof Fons van de Vijver (Tilburg University) and Prof Ian Rothmann (North-West University, Potchefstroom Campus).

Meiring Family, Deon, Ilse, Chloé, Kylie
Artwork created by Chloé & Kylie Meiring